THE UNIVERSITY
of EDINBURGH

Population and Functional Genomic Analysis of *Lawsonia intracellularis*

Rebecca Ji Bengtsson

Thesis presented for the degree of Doctor of Philosophy

The University of Edinburgh

2018

I declare that work contained within this thesis has been composed my me and is entirely my own work, unless otherwise stated. This thesis has not been submitted for any other degree or professional qualification.

Rebecca Ji Bengtsson December 2018

Acknowledgments

Firstly, I would like to thank my supervisors Professor Ross Fitzgerald and Dr Tahar Ait-Ali for their continuous guidance and support. Thank you both for pushing me to always do my best, providing me with encouragement and motivation throughout my PhD. I would also like to say a special thanks to Professor Alan Archibald and Dr Eleanor Watson for their help and input on my work.

My sincere thanks to the current and past members of LBEP especially Bryan, Amy, Gonzalo, Emily, Manouk, Joana and Rodrigo. I am most grateful for all for the help they have provided me over the years, taking time off from their own research to listen to me practice my talk before a presentation and for all the helpful discussions on my work. I feel truly blessed to have had worked with a group of such wonderful and talented lab members.

In addition, my heartfelt thanks to all members of the Archibald, Gally, Morrison and Stevens group for all the fun and laughter we have had during the past four years. Thank you all for keeping me sane through my PhD and made my time in the Roslin Institute so delightful.

Finally, I would like to thank my family and friends for their endless love and support. Thank you mum, dad, Ida and Emil for always believing in me, this would not have been possible without you all. My girls Vicki, Omayma, Linda, Pip, Wiki and Kate, thank you all for always being there for me and cheering me on till the finish line.

Table of Contents

Table of Contents **7**

List of Figures **11**

List of Tables **13**

Abbreviations **15**

Abstract **19**

Lay Summary **21**

CHAPTER 1 25

1.1 Metagenomics in microbiology **27**

1.1.1 Expanding the tree of life 27

1.1.2 Pathogen detection in infectious disease 28

1.2 Bioinformatics methods in metagenomic analysis **30**

1.2.1 Metagenome assembly 30

1.2.2 Binning contigs 36

1.2.3 Taxonomic classification 38

1.3 Bacterial genomics **39**

1.3.1 Clonality in bacterial population 39

1.3.2 Recombination and horizontal gene transfer 40

1.3.3 Bacterial population structure 41

1.3.4 Tools for investigating bacterial genetic diversity and population structure 43

1.3.5 Bacterial genomics in infectious diseases 44

1.4 Evolution and diversity of obligate intracellular bacteria **48**

1.4.1 Genome reduction in obligate intracellular bacteria 48

1.4.2 Mutualistic versus parasitic association 49

1.4.3 Horizontal gene transfer in obligate intracellular pathogens 50

1.5 Introduction to *Lawsonia intracellularis* **51**

1.5.1 Clinical characteristics of proliferative enteropathy 51

1.5.2 Pathogenesis of *L. intracellularis* infection 53
1.5.3 Transmission of *L. intracellularis* 56
1.5.4 Host association of *L. intracellularis* 57
1.5.5 Genomic features of *L. intracellularis* 60
1.5.6 Clinical and economical significance of *L. intracellularis* 62
1.5.7 Control of *L. intracellularis* 62

1.6 Background summary 64

1.7 Project Aims 65

CHAPTER 2 67

2.1 Introduction 69

2.2 Aims 70

2.3 Materials and Methods 71
2.3.1 Bacterial strains, DNA extraction and microbial DNA enrichment 71
2.3.2 Quantitative Polymerase Chain Reaction 71
2.3.3 Genome sequencing and sequence processing 72
2.3.4 Taxonomic classification and host reads removal 73
2.3.5 Genome assembly and annotation 74
2.3.6 Genome assembly quality assessments 75

2.4 Results 77
2.4.1 Development of a quantitative assay to measure *L. intracellularis* DNA from complex samples 77
2.4.2 Sequencing analysis of clinical samples 81
2.4.3 Microbial DNA enrichment maximised sequencing depth of coverage from tissue samples 87
2.4.4 Assessment of metagenomic assemblers on *L. intracellularis* draft genome reconstruction 95
2.4.5 Unsupervised binning unable to recover *L. intracellularis* plasmid 1 and plasmid 2 101
2.4.6 A reference-guided *de novo* assembly approach improves draft genome reconstruction 107

2.5 Discussion 114

CHAPTER 3 121

3.1 Introduction **123**

3.2 Aims **124**

3.3 Materials and Methods **125**

3.3.1 Whole genome sequence alignment and phylogenetic inference 125

3.3.2 Variant calling 125

3.3.3 Recombination analysis 126

3.3.4 Pangenome analysis 127

3.3.5 Positive selection analysis 127

3.3.6 Dating analysis with BEAST 128

3.4 Results **129**

3.4.1 Population genomic structure of *L. intracellularis* reveals host restricted genetic diversity 129

3.4.2 Porcine derived *L. intracellularis* isolates revealed limited genetic diversity 132

3.4.3 Pan-genome analysis revealed *L. intracellularis* isolates have highly conserved gene content variation 138

3.4.4 Multiple high SNP density regions across the *L. intracellularis* genome contribute to genetic variation between the porcine and equine clades 143

3.4.5 Genes involved in energy production, signal transduction and membrane biogenesis are divergent among *L. intracellularis* isolates 150

3.4.6 Putative genes involved in host cell invasion and stress response display signatures of positive selection 154

3.4.7 Estimation of the timeframe for evolution of *L. intracellularis* 156

3.5 Discussion **159**

CHAPTER 4 165

4.1 Introduction **167**

4.2 Aims **168**

4.3 Materials and Methods **169**

4.3.1 Sequence analysis of LatB 169

4.3.2 Clinical and cell cultured *L. intracellularis* samples used in the current study 169

4.3.3 Genomic DNA extraction 169

4.3.4 Sequencing of *latB* 170

4.3.5 Synthetic LatB peptide and anti-LatB IgG antibody 171

4.3.6 Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis and Coomassie Blue Staining 171
4.3.7 Western blot analysis 172
4.3.8 Immunohistochemistry 173
4.3.9 *latB* cloning for recombinant protein expression 174
4.3.10 Recombinant Protein Expression 177
4.3.11 Native purification of recombinant Glutathione S-transferase (GST)-tagged proteins 177
4.3.12 Preparation of *L. intracellularis* Enterisol® Ileitis surface antigens 178
4.3.13 Enzyme-Linked Immunosorbent Assay 179

4.4 Results 180
4.4.1 Structural prediction of LatB 180
4.4.2 Variation in the length of *latB* were observed among *L. intracellularis* isolates 184
4.4.3 LatB is expressed during *L. intracellularis* infection 190
4.4.4 Generation of recombinant LatB proteins for immunological investigations 197
4.4.5 Low levels of immunological response to LatB during *L. intracellularis* infection 201

4.5 Discussion 206

CHAPTER 5 213

References 227

Appendix. Supplementary Tables 249

List of Figures

Chapter 1

Figure 1.1 k-mer length variation and DBG construction.....33
Figure 1.2 Impact of single nucleotide variation on DBG construction. 35
Figure 1.3 Flow chart outlining the pipeline for unsupervised binning. 37
Figure 1.4 Illustration of different evolutionary scenarios operating within a bacterial population from T1 – T7 time points. 44
Figure 1.5 Bacterial genomics in infectious diseases.....47
Figure 1.6 Schematic diagram displaying cross-species infection of *L. intracellularis*. 59

Chapter 2

Figure 2.1 PCR amplification of *L. intracellularis* 16S rRNA and *aspA*.....79
Figure 2.2 Correlation between *L. intracellularis aspA* copy numbers in sequencing library and sequencing outcome.....84
Figure 2.3 Assessment of microbial enrichment to enhance *L. intracellularis* sequencing from ileal DNA. 91
Figure 2.4 Assessment of microbial enrichment to enhance *L. intracellularis* sequencing from faecal DNA.....93
Figure 2.5 Detection of *L. intracellularis* DNA from different enrichment fractions...94
Figure 2.6 Quality assessment plot of *L. intracellularis* genome bins.....103
Figure 2.7 *L. intracellularis* genome bins. 106
Figure 2.8 Flow chart outlining the pipeline for reference-guided and reference independent de novo assembly.....109

Chapter 3

Figure 3.1 Unrooted ML tree generated based on core genome alignments of 28 *L. intracellularis* isolate. 131
Figure 3.2 ML Phylogeny of porcine *L. intracellularis* isolates across Europe, Central and South America.....134
Figure 3.3 Diagram displaying the distribution of SNPs across the genomes of *L. intracellularis* porcine isolates.....136
Figure 3.4 Pan-genome analysis of 28 *L. intracellularis* isolates.....141

Figure 3.5 Comparison of the putative ZIP family metal transporter locus. Comparison of the gene locus among three isolates. 142
Figure 3.6 Gubbins recombination analysis output. 146
Figure 3.7 Distribution of regions with elevated SNP density across *L. intracellularis* genomes. 149
Figure 3.8 Root-to-tip regression panel for the 21 isolates. 157

Chapter 4

Figure 4.1 Schematic diagram showing the domain structure of LatB from PHE/MN1-00 strain. 183
Figure 4.2 Variation of *latB* passenger domains among nine clinical *L. intracellularis* isolates. 187
Figure 4.3 Western blot analysis of LatB passenger domain in *L. intracellularis* Enterisol® Ileitis strain cell lysate. 192
Figure 4.4 IHC of LatB in ileum tissue sections of pigs experimentally challenged with *L. intracellularis* LR189 strain. 194
Figure 4.5 IHC of LatB in ileum tissue sections of pigs experimentally challenged with *L. intracellularis* LR189 strain. 196
Figure 4.6 Western blot analysis of recombinant LatB (rLatB) expression 200
Figure 4.7 Pigs naturally infected with *L. intracellularis* produce antibodies against Enterisol® Ileitis surface antigens. 203
Figure 4.8 Low levels of serum antibodies specific for rLatB in pigs infected with *L. intracellularis* 205

Chapter 5

Figure 5.1 Metagenomic sequencing of clinical samples for investigation of population and functional genomics of *L. intracellularis*. 225

List of Tables

Chapter 2

Table 2.1 Primers and gBlock gene fragment sequence employed in the current study. 76
Table 2.2 Sequencing statistics of samples processed after 16S rRNA based quantification. 80
Table 2.3 Sequencing statistics of faecal DNA samples. 85
Table 2.4 Sequencing statistics of enriched and non-enriched DNA samples..... 89
Table 2.5 Comparison of assembly quality using different metagenomic assembler. 97
Table 2.6 Performance comparison of metagenomic assemblers..... 99
Table 2.7 Table describing completeness and contamination of each *L. intracellularis* genome bin. 104
Table 2.8 Comparison of assembly statistics between reference-guided and reference independent assembly. 110
Table 2.9 *L. intracellularis* draft genomes assembled using reference-guided approach with metaSPAdes. 112

Chapter 3

Table 3.1 *L. intracellularis* isolates used for comparative analysis. 130
Table 3.2 Nonsense mutations identified among porcine associated *L. intracellularis* isolates. 137
Table 3.3 Base substitution statistics detected by Gubbins. 147
Table 3.4 Genes with highly divergent protein sequences between isolates from equine and porcine clade. 152
Table 3.5 Genes with signatures of positive selection in *L. intracellularis*. 155
Table 3.6 Estimated evolutionary rate and divergence times for the tMRCA of *L. intracellularis* porcine isolates under different demographic models in Beast2..... 158

Chapter 4

Table 4.1 Primers used for *latB* cloning for recombinant protein expression..... 176
Table 4.2 Variation of latB passenger domain in clinical isolates. 188
Table 4.3 Variation of LatB passenger domain in the cell passaged isolates. 189

Abbreviations

Abbreviations	Full Description
h-path	Hub-path
aa	Amino acid
AAI	Amino acid identity
AC	Autochaperone
ANI	Average nucleotide identity
AT	Autotransporter
BEAST	Bayesian evolutionary analysis by sampling trees
BLAST	Basic local alignment search tool
CDS	Coding sequence
COG	Clusters of orthologous groups
DBG	De Brujin graph
dN	Non-synonymous mutation
DNA	Deoxyribnucleic acid
DPI	Days post infection
dS	Synonymous mutation
ELISA	Enzyme-linked immunosorbent assay
GC	Guanine-cytosine
GI	Gastrointestinal
GSC	Genomic Standards Consortium
GTR	General time reversible

h	Hour
HGT	Horizontal gene transfer
Hi-C	Genome-wide chromatin interaction
Indel	Insertions and deletions
kb	Kilobase
kDa	Kilodalton
LatB	*Lawsonia* autotransporter B
LCA	Lowest common ancestor
MAG	Metagenome-assembled genome
Mb	Megabase
MBD2	Methyl-CpG-binding domain protein 2
MIMAG	Minimum Information about a Metagenome-Assembled Genome
min	Minute
ML	Maximum likelihood
ml	Mililitres
MLEE	Multilcous enzyme electrophoresis
MLST	Multilcous sequence typing
NCBI	National centre of biotechnology information
N_e	Effective population size
NGS	Next generation sequencing
OM	Outer membrane
PDBG	Paired de Bruijn graph
PE	Proliferative enteropathy
PFGE	pulsed field gel electrophoresis
PHE	Porcine haemorrhagic enteropathy

PIA	Porcine intestinal adenomatosis
RAxML	Randomized Accelerated Maximum Likelihood
rRNA	Ribomosomal RNA
SAG	Single-cell assembled genome
SDBG	Succinct de Bruijn graph
SNP	Single-nucleotid polymorphism
TNF	Tetra-nucleotide frequency
VNTR	Variable number tandem repeat
WGS	Whole genome sequencing

Abstract

The Gram-negative, obligate intracellular bacterium *Lawsonia intracellularis* is the aetiological agent of a non-zoonotic enteric disease known as proliferative enteropathy (PE). PE has been detected in wide range of mammalian species, including wild and domestic animals. The disease is most prevalent in pig herds where it is endemic, with both subclinical and clinical infections. Porcine PE was first described in 1931 and four different form of the disease are recognised, ranging from mild to severe clinical manifestations. Over recent years rising outbreaks of equine PE in foals have been reported and cross-species experimental infections have demonstrated host-specificity for infection. Although the dynamics of *L. intracellularis* infection have been examined, there is a lack of understanding population genetic structure and the basis for pathogenesis due to difficulties with culturing obligate intracellular organisms.

In this study, we developed a pipeline to obtain high quality whole genome sequences of *L. intracellularis* through direct sequencing of clinical samples from field outbreaks of porcine and equine PE, and cell passaged samples without the requirement of cultivation, resulting in the generation of 25 draft genome assemblies. To explore the genetic diversity of *L. intracellularis*, we performed comparative analysis which revealed a clonal population structure among porcine derived isolates showing very limited diversity, indicative of emergence from a recent clonal expansion. Our analysis revealed that infections among different hosts are caused by genetically distinct *L. intracellularis* sub-types. Finally, we focussed on a core gene encoding a putative autotransporter, termed *Lawsonia* autotransporter B (*latB*), that exhibited genetic variation in different strains, and immunofluorescent staining for LatB revealed the

putative autotransporter was expressed during infection and was associated with the bacterial outer membrane.

Overall, these studies have provided insights into the genetic diversity and evolutionary origin of L. intracellularis, enhancing our understanding of the population biology of this important animal pathogen.

Lay Summary

The bacterium *Lawsonia intracellularis* is the cause of a common gastrointestinal disease known as proliferative enteropathy (PE) that is widespread across the world, affecting a wide range of mammalian species, most notably pigs. Different forms of the disease have been described in pigs, which vary in clinical severity. The mild form is characterised by diarrhoea, weight loss and anorexia whereas the severe form is characterised by haemorrhagic diarrhoea and sudden death. In addition, a high percentage of disease episodes in pigs are subclinical, whereby animals do not present clinical signs or symptoms but may continue to lose weight.

Currently, the underlying cause of variation in disease severity is unknown, as we have very limited understanding of the pathogen genome evolution and the genetic basis of *L. intracellularis* pathogenesis. This is mainly due to difficulties with culture in laboratory conditions hindering research into the biology of the organism. In the current study, we utilised state-of-the-art sequencing technologies to obtain genome sequences of *L. intracellularis* through direct sequencing of clinical samples from animals with PE. Analysis of the genome sequences showed that *L. intracellularis* isolates derived from pigs are highly similar with low levels of genetic diversity among the population, which suggests a recent clonal expansion from a single strain. Comparison of isolates derived from pigs and horses revealed that isolates derived from different hosts are genetically distinct, and may represent different subtypes of *L. intracellularis* that are species-specific. Finally, we identified length variation in a gene, coding for a putative surface protein among *L. intracellularis* isolates, in which variation appeared to be strain specific and not associated with virulence.

Proliferative enteropathy has a significant impact on health, welfare and animal production, imposing a huge economic burden on the pig farming industry. The current analysis has increased our understanding of the biology and genome evolution of *L. intracellularis* which will inform novel strategies for the control of proliferative enteropathy.

Chapter 1

General introduction

1.1 Metagenomics in microbiology

Metagenomics refers to the study of all genetic materials recovered from a community within an environmental sample. The earliest metagenomic studies focused on sequencing of the 16S ribosomal RNA (rRNA) gene, which helped uncover the extent of microbial diversity that exists in various environments such as marine sediments, ocean surfaces, human and animal guts (Gray and Herwig, 1996, Biers et al., 2009, Gill et al., 2006). Metagenomic studies have now expanded to whole genome sequencing which provides insights beyond phylogenetic diversity, facilitating the exploration of metabolic and functional diversity of microbial communities (Rinke et al., 2013, Brown et al., 2015, Anantharaman et al., 2016). Currently, metagenomic analysis is being applied in all areas of microbiology.

1.1.1 Expanding the tree of life

Microbial genomes currently deposited in public repositories are heavily biased towards organisms with medical importance and those that can be cultured in laboratory conditions but do not represent the naturally occurring diversity (Solden et al., 2016). The culture-independent approach of metagenomic sequencing has enabled recovery of microbial genomes from communities without cultivation (Parks et al., 2017, Stewart et al., 2018). Consequently, the numbers of single-cell assembled genomes (SAGs) and metagenome assembled genomes (MAGs) registered in the Genome Online Database (GOLD) has risen significantly (Reddy et al., 2014).

A newly constructed tree of life built with genomes available from public databases and over 1,000 MAGs recovered from a variety of environments, dramatically

expanded the diversity within the Bacteria, Archaea and Eukarya domains (Hug et al., 2016). The tree revealed a large number of major lineages without representatives derived from laboratory culture, further highlighting the lack of true biodiversity represented by sequences in public databases (Hug et al., 2016).

1.1.2 Pathogen detection in infectious disease

The emerging field of metagenomics has also proven a useful tool in clinical settings for the detection and investigation of pathogens in infectious disease (Loman et al., 2013, Wilson et al., 2015, Doan et al., 2016, Pendleton et al., 2017). Currently, surveillance and investigation of infectious diseases predominantly relies on conventional laboratory testing assays including microscopy, immunoassays, PCR-based and culture-based analyses. These processes are time consuming and inefficient. Furthermore, analysis often requires *a priori* knowledge of the pathogen of interest. It has been reported that through the use of conventional assays, in up to 40% of human diarrhoea and 60% of encephalitis cases, no etiologic agents can be identified (Finkbeiner et al., 2008, Ambrose et al., 2011). The untargeted approach of metagenomic sequencing enables the majority of microbial pathogens in a sample to be identified in a single assay. The additional sequence data provided by metagenomic sequencing may also facilitate prediction of virulence through presence of drug resistance genes or toxins, phylogenetic reconstruction for outbreak investigation and pathogen evolution studies (Benjak et al., 2018). Furthermore, the application of metagenomic sequencing for pathogen detection may offer rapid diagnosis and management of infectious diseases involving fastidious organisms (Hasman et al., 2013), such as *Mycobacterium tuberculosis* in which traditional culture-based testing may take several weeks or even months. Whereas, through

metagenomic sequencing, the entire process of DNA extraction from sputum samples to sequencing and analysis can be performed within a few days (Doughty et al., 2014). Moreover, multiple causative agents in mixed infections can be identified through metagenomic sequencing (Kujiraoka et al., 2017, Li et al., 2018), and enable investigation of complex diseases involving microbiome dysbiosis, such as type 2 diabetes and inflammatory bowel disease (Qin et al., 2012, Greenblum et al., 2012).

Although integration of metagenomics into clinical laboratories for diagnostic use has great advantages, it does however come with limitations, including low coverage of pathogen genomes which can impact on diagnostic sensitivity and accuracy (Schlaberg et al., 2017). Nevertheless, with the continuing decrease in sequencing costs and advances in high-throughput sequencing technologies addressing the current challenges, metagenomics has great potential for future use in clinical contexts.

1.2 Bioinformatics methods in metagenomic analysis

In conventional shotgun sequencing of a single microbial isolate, all sequenced reads are typically assumed to be derived from clones of the same genome and coverage along the genome will be approximately uniform. Conventional *de novo* assemblers therefore utilise these properties to correct for sequencing errors (Zerbino and Birney, 2008, Simpson et al., 2009), identify sequence variants and resolve repeat regions (Iqbal et al., 2012). In contrast, metagenomic sequence data contains a complex mixture of reads from a wide variety of microorganisms that vary in abundance which presents unique challenges for sequence assembly and analysis (Quince et al., 2017). However, with the rise in metagenomic sequencing follows a suite of bioinformatic approaches to overcome these challenges.

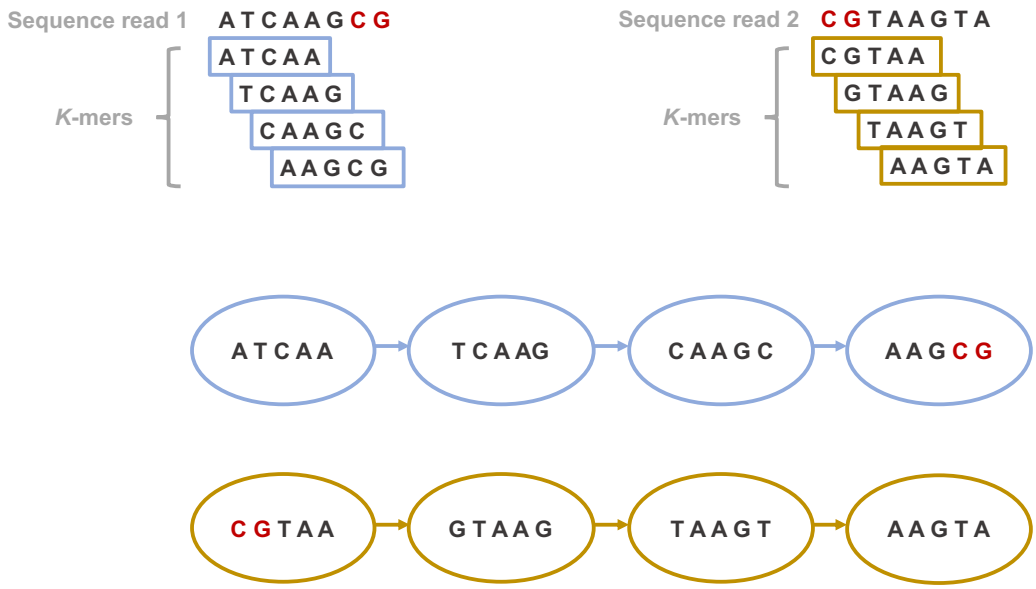
1.2.1 Metagenome assembly

Most existing assemblers for metagenomic shotgun sequencing use the de Bruijn graph (DBG) approach. This involves breaking reads down into overlapping substrings of fixed length, known as *k*-mers and organising them into a DBG of overlapping *k*-mers with *k*-1 matching bases (Pevzner et al., 2001). Contigs are then assembled by traversing through a path across the graph (Pevzner et al., 2001). Selection of *k*-mer length is an important parameter for DBG construction, by which a trade-off between a graph that is less tangled but more fragmented, versus a graph that is more contiguous and more tangled exists (Figure 1.1). The presence of repeats, sequencing errors and chimeric reads may complicate DBG construction by introducing multiple path ambiguities in the graph resulting in genome mis-assemblies and fragmentation.

Sequencing coverage of genomes is highly variable within metagenomic data. Since genome coverage corresponds to genome abundance, varying levels of different species within the sampled community will result in a highly non-uniform sequencing coverage of genomes. Selection of a short *k*-mer length may recover low abundant genomes but increase the chance of mis-assemblies, while a longer *k*-mer length will provide more accurate contigs but bias towards highly-abundant genomes (Figure 1.1) (Vollmers et al., 2017). Assemblers developed for metagenome assembly address this issue by applying a multi *k*-mer approach which iterates from short to long k-mer length during assembly to account for reconstruction of low and highly abundant species (Vollmers et al., 2017).

Strain resolution is another challenge for metagenome assembly, as within a microbial community, there may be a mixture of strains of varying abundance that can cause problems for DBG construction similar to those presented by sequencing errors (Figure 1.2) (Olson et al., 2017). Thus, metagenomic specific assembly algorithms are unable to assemble every strain variant in the community but will capture the dominant haplotype and provide best representation of each species (Olson et al., 2017).

(A)


Sequence read 1
A T C A A G C G
K-mers
A T C A A
T C A A G
C A A G C
A A G C G
Sequence read 2
C G T A A G T A
K-mers
C G T A A
G T A A G
T A A G T
A A G T A
A T C A A
T C A A G
C A A G C
A A G C G
C G T A A
G T A A G
T A A G T
A A G T A

(B)

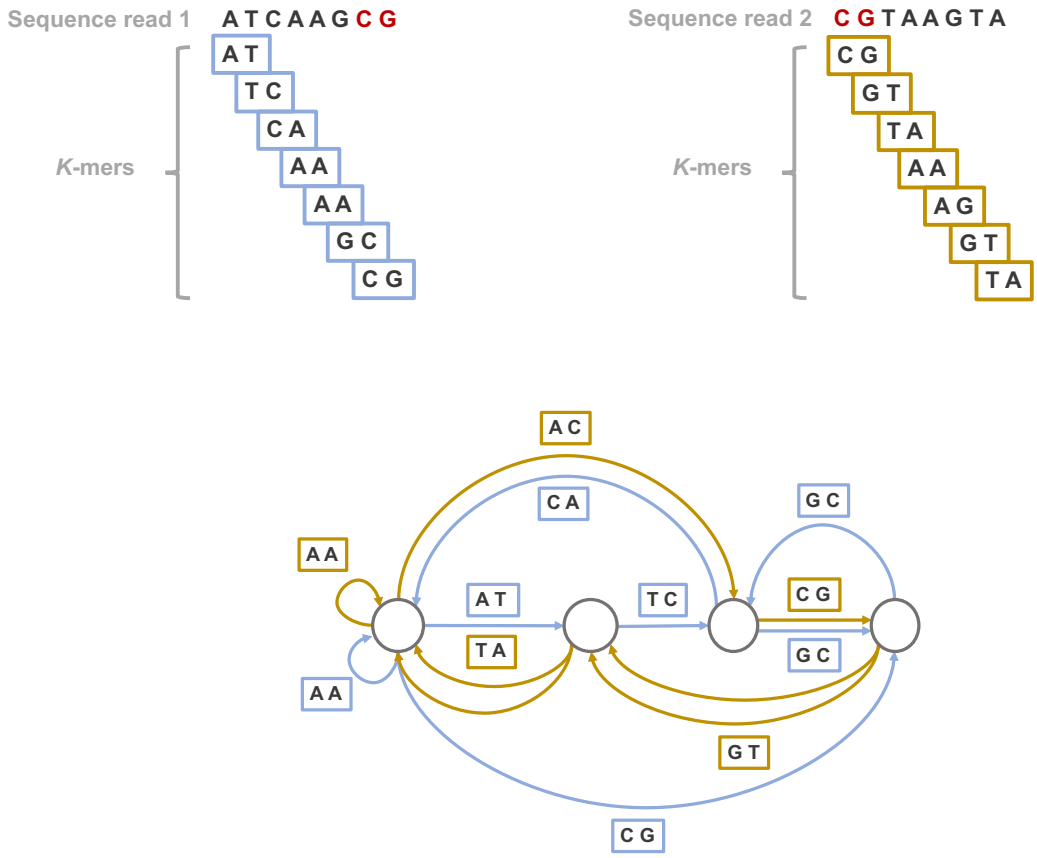
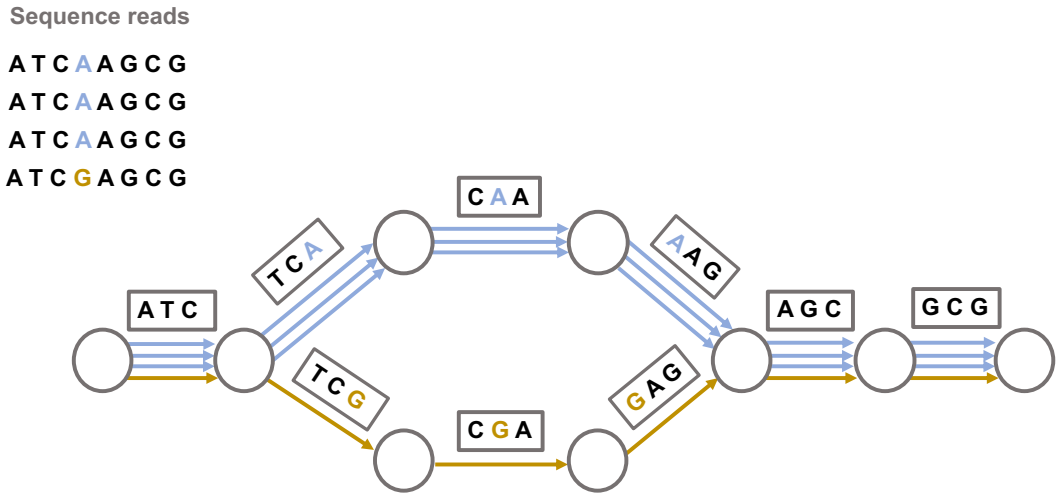

Sequence read 1
A T C A A G C G
K-mers
A T
T C
C A
A A
A A
G C
C G
Sequence read 2
C G T A A G T A
K-mers
C G
G T
T A
A A
A G
G T
T A
A C
C A
G C
A A
A T
T C
C G
T A
G C
A A
G T
C G

Figure 1.1 k-mer length variation and DBG construction. (A) Setting a larger k-mer length of 5 may increase the abundance of unique k-mers, this resolves the problem of short repeats, producing a less tangled graph. The two adjacent reads overlap one another by two base pair (marked in blue), but because k-mer length is set to 5 and the two adjacent reads do not sufficiently overlap by the required length of 4 (k-1mer), the path is broken. (B) Setting a smaller k-mer length of 2 will provide sufficient overlapping of the two reads and a contiguous graph. However, this will increase the abundance of repetitive k-mers (AA, AG and CG with a frequency of f=2) in the data and introduce more branching and ambiguities in the graph. Image adapted from (Vollmers et al., 2017).

(A)


Sequence reads
ATCAAGCG
ATCAAGCG
ATCAAGCG
ATCGAGCG
ATC
TCA
TCG
CAA
CGA
AAG
GAG
AGC
GCG

(B)

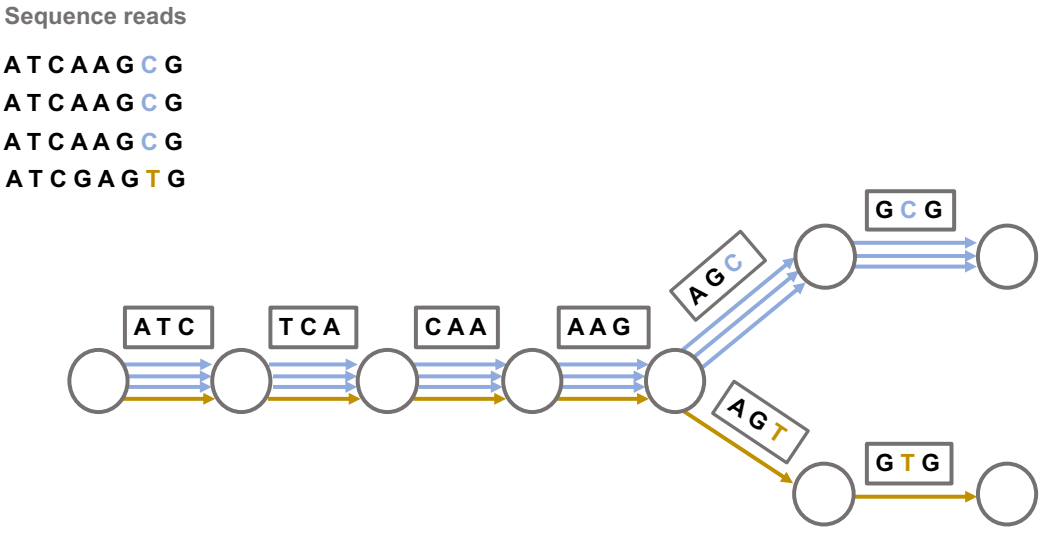

Sequence reads
ATCAAGCG
ATCAAGCG
ATCAAGCG
ATCGAGTG
ATC
TCA
CAA
AAG
AGC
AGT
GCG
GTG

Figure 1.2 Impact of single nucleotide variation on DBG construction. (A) A miscalled base as a result of sequencing error in the middle of a read will lead to a "bubble" structure. (B) A miscalled base occurring near the beginning or end of a read will lead to "tip" structure. In both case sequencing error introduce h-paths along the DBG by which only one path is correct (green). In both cases, the erroneous k-mer and incorrect path (red) is occurring at lower frequency, these can be filtered by using a defined coverage cut-off and removed from the data based on low coverage. Image adapted from (Vollmers et al., 2017).

1.2.2 Binning contigs

Metagenomic assemblies produce fragmented contigs derived from genomes of hundreds to thousands of species within a community. The process of contig 'binning' involves grouping contigs which are derived from the same genome together in an imaginary bin (Tyson et al., 2004). Binning can be performed either supervised or unsupervised. Supervised binning involves the use of databases to taxonomically classify or group contigs based on homology with known reference genomes (Quince et al., 2017). Unsupervised binning involves clustering contigs from the same genome together based on various shared properties (Figure 1.3). Depending on the algorithm used this may include tetranucleotide frequencies, coverage profiles and intra-genome GC content (Dick et al., 2009). Clustering of contigs via unsupervised binning has the advantage of not requiring reference genomes. However, it may lack the discriminatory power to resolve strain-level heterogeneity within a complex community (Doughty et al., 2014).

Binning was first applied to shotgun metagenomic sequencing data from a natural acidophilic biofilm from which two near-complete microbial genomes were recovered (Tyson et al., 2004). Since then, this approach has been adapted in large-scale recovery of 8,000 and 913 draft quality genomes from public metagenomes and cow rumen samples, respectively (Parks et al., 2017, Stewart et al., 2018).

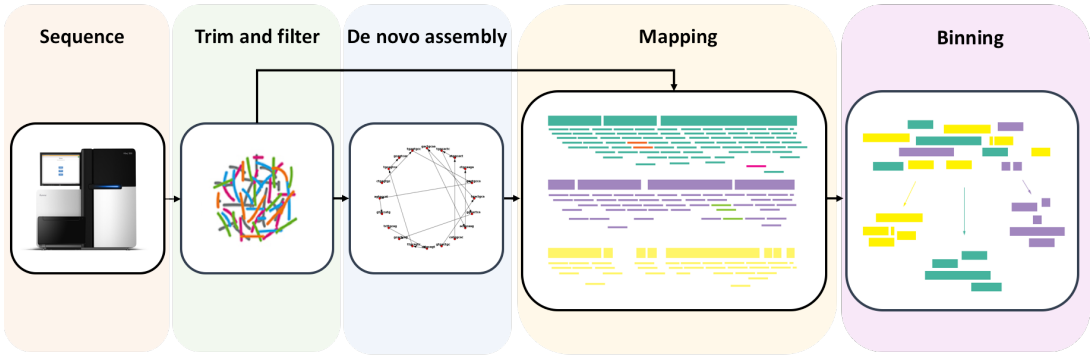


Figure 1.3 Flow chart outlining the pipeline for unsupervised binning. Following metagenomic assembly, quality filtered and adapter trimmed reads are mapped back to the assembled contigs to acquire coverage profiles. Sequence composition of the contigs, such as tetranucleotide frequencies and GC content are calculated. Contigs originating from the same genome share similar coverage profile and sequence composition, and are grouped together in a genome bin accordingly.

1.2.3 Taxonomic classification

Taxonomic classification is commonly performed using a database of marker genes that are clade-specific single-copy core genes and are ubiquitous across bacterial and archaeal genomes to assign the taxon of binned draft genomes. In addition, this approach is often used to determine the quality of MAGs by evaluating the presence or absence of these marker genes, in turn assessing the level of completeness and contamination in each genome bin (Parks et al., 2015).

Taxonomic profiling of metagenomic data may also be performed without assembly, by mapping metagenomic reads against databases containing phylogenetically distinctive marker sequences (Quince et al., 2017). This approach avoids the problems associated with metagenome assembly as it provides strain-level resolution for population genomics and epidemiological studies, and allows classification of low-abundance genomes which may not be assembled due to lack of coverage (Truong et al., 2017, Luo et al., 2015, Truong et al., 2015).

1.3 Bacterial genomics

Bacterial population structure is shaped by various evolutionary processes and forces. By gaining an understanding of these processes will help us interpret the response of bacterial population to various selection pressures, improve our understanding of pathogen biology and help us design interventions for control of the disease.

1.3.1 Clonality in bacterial population

Bacterial genomes evolve through the processes of mutation and recombination. The former occurs when errors are made during DNA replication or DNA damage repair, generating point mutations (Spratt and Maiden, 1999). Because bacteria reproduce asexually through binary fission, genetic variation generated by mutation will be vertically inherited by all progeny bacteria. Consequently, the distribution of genetic variation among a vertically descended bacterial population will exhibit linkage disequilibrium (Spratt and Maiden, 1999).

Clonality of bacterial lineages can be disrupted by sex, which involves inheritance of exogenous genetic material through recombination (Spratt, 2004). In the total absence of recombination, genetic variation arises through mutation and leads to a clonal population (Spratt and Maiden, 1999). However, truly clonal bacterial species in which no recombination events have occurred during their evolutionary history are rare. The occurrence of linkage disequilibrium can be attributed either by periodic selection or genetic drift, or both (Cohan and Perry, 2007). Genetic drift results in the reduction of sequence diversity due to random elimination of lineages, by which the effect is much greater in small bacterial populations, compared to large populations

(Cohan and Perry, 2007). Periodic selection occurs when an adaptive mutation has been acquired that confers competitive advantage within a particular niche has been acquired, followed by natural selection that drives the expansion of the adapted clone to outcompete others in the population and increase in frequency, resulting in a genome-wide sweep (Cohan and Perry, 2007). This will lead to occasional epidemic clonal population structure, in which the emergence of a successful clonal lineage consequently increased its capacity to cause disease and dominate the population (Figure 1.4) (Spratt and Maiden, 1999).

1.3.2 Recombination and horizontal gene transfer

Sexual events in bacteria involves inheritance of exogenous genetic materials through horizontal gene transfer (HGT) by mechanisms of transformation, transduction and conjugation (Narra and Ochman, 2006). Through recombination, bacteria may incorporate exogenous genetic materials into its genome to acquire novel combinations of alleles (Narra and Ochman, 2006). Thus, the distribution of genetic variation among a bacterial population will exhibit linkage equilibrium (Spratt and Maiden, 1999).

As previously mentioned, recombination can disrupt clonality and increase genetic variation in the population, by shuffling of alleles. Furthermore, recombination may provide evolutionary benefits through introduction of favourable alleles or sequences to the genome, which may facilitate rapid adaptation within an ecological niche and in some cases, enable the microorganism to occupy a new niche (Narra and Ochman, 2006, Cohan and Perry, 2007). The transfer of antibiotic resistance gene through HGT is one of the most prominent examples, by which acquisition of the virulence factor

has enhanced and facilitated epidemic emergence of various pathogens (Baker et al., 2018, Navon-Venezia et al., 2017, Price et al., 2012).

1.3.3 Bacterial population structure

The rate of recombination varies greatly among bacterial species (Narra and Ochman, 2006). While clonal pathogens such as *Mycobacterium tuberculosis*, in which recombination is rare and exhibits a clonal population structure with low sequence diversity (Gagneux, 2018). Pathogens such as *Helicobacter pylori*, undergo frequent intraspecific recombination many orders of magnitude more frequently than mutation and exhibit extensive population genetic diversity (Suerbaum and Achtman, 2004).

Bacterial population structure is shaped by the relative rates of mutation and recombination, which is influenced by natural selection, the ecological niche and effective population size (N_e) of the bacterial species (Achtman and Wagner, 2008). A clonal bacterial population exhibiting linkage disequilibrium, may have resulted from diversity-reducing events such as periodic selection or genetic drift (Figure 1.4) (Cohan and Perry, 2007). This effect can be eliminated by recombination, whereby genetic variations are introduced (Spratt and Maiden, 1999). However, the rate of recombination will be influenced by the availability of novel genes within the ecological niche, and the ability for the bacteria to retain the exogenous DNA, which is in turn determined by the ecological selection pressure (Shapiro, 2016).

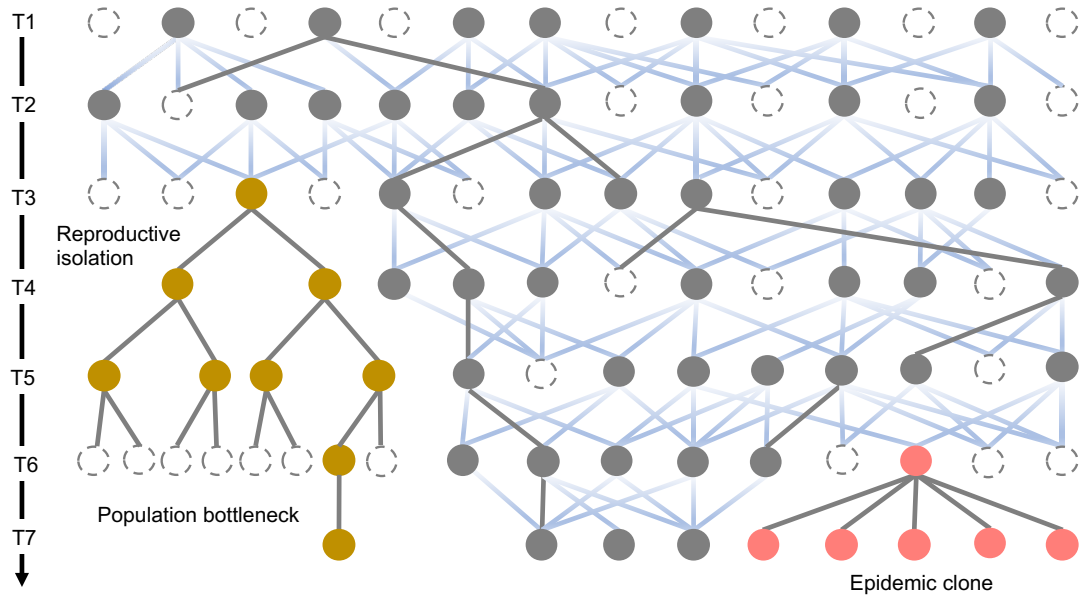


Figure 1.4 Illustration of different evolutionary scenarios operating within a bacterial population from T1 – T7 time points. Grey dots represent bacterial cells that contribute offspring to the next generation and white dots represents members that will not. Over generations members will can reproduce will outnumber those who cannot. Blue gradient lines represent horizontal transmission of genetic materials and grey lines represent vertical transmission events. At T3, events which may lead to emergence of a clonal subpopulation (mustard dots) that becomes reproductively isolated and can only transmit genetic material through vertical descent. At T6 – T7, population bottleneck will result in a further decrease of the population genetic diversity. The pink dot at T6 represents acquisition of beneficial mutation or genes that increase the isolates capacity to cause disease. At T7 this epidemic clone has spread rapidly and becomes over represented within the population. Image adapted from (Spratt and Maiden, 1999).

1.3.4 Tools for investigating bacterial genetic diversity and population structure

Traditional molecular typing methods for investigating bacterial population structure and phylogeny includes multilocus enzyme electrophoresis (MLEE), multilocus variable number tandem repeat (MLVA) (Schouls et al., 2009), multilcous sequence typing (MLST) (Maiden et al., 1998) and pulsed-field gel electrophoresis (PFGE) (Singh et al., 1999). While MLEE, MLVA and MLST rely on sequence analysis from a few genes or loci (Singh et al., 2006). PFGE provides a measurement of differences spanning the entire genome and long been considered the gold standard approach for molecular epidemiological investigation (Swaminathan et al., 2001). However, reproducibility among different laboratories can be difficult to achieve and interpretations can be discordant (Goering, 2010). Among these methods, MLST is considered the more robust approach and involves sequencing multiple genes (usually six to eight) that are under low selective pressure and distributed around the chromosome (Singh et al., 2006). Due to its ability to unambiguously differentiate between clones within pathogen species and resolve ancestral relationships among isolates, this method has been one of the most popular and most widely used approach for investigating bacterial genetic diversity (Bentley and Parkhill, 2015). However, a study comparing phylogenies inferred from MLST and whole genome data of 10 bacterial species were incongruent and showed that MLST failed to fully represent true bacterial genome phylogenies (Tsang et al., 2017). Furthermore, by using only a small portion of the genome this method lacks sufficient discriminatory power for distinguishing among highly clonal pathogens, closely related outbreak strains or clonally dominant endemic strains (Brodrick et al., 2016, Achtman, 2008).

The introduction of high throughput sequencing (HTS) has enabled quick, accurate and cheap whole genome sequencing (WGS) of microbial pathogens, able to discriminate between near-identical strains and detect rare genetic variants (Bentley and Parkhill, 2015). In turn, improved resolution for population genomic investigations and facilitates epidemiological investigation of pathogens during endemic outbreak (McAdam et al., 2014). For assessment of clonality, WGS has advantage over MLST of displaying the full scale of recombination events occurring along the bacterial genome (Bobay et al., 2015). Since in most bacterial species, recombination often impacts only small regions of the genome, MLST may not capture this (Bobay et al., 2015). Indeed, WGS has demonstrated that the extent of recombination in clonal pathogens such as *Escherichia coli* and *Staphylococcus aureus* is higher than previously estimated (Mau et al., 2006, Didelot et al., 2012, Takuno et al., 2011). Furthermore, WGS is able to provide information beyond the core genome and provides a complete picture of the gene repertoire, enable identification of accessory mobile elements such as bacteriophage, plasmids and pathogenicity islands, known to carry virulence genes (Bentley and Parkhill, 2015).

1.3.5 Bacterial genomics in infectious diseases

The genome sequence of a bacterium encodes all the genes necessary for the organism's lifestyle, and its analysis provides unique insights into the physiology and evolution of the species. Comparative analyses can reveal differences in genomic compositions and organisation among bacterial pathogens, this information coupled with clinical, epidemiological and temporal data is a powerful approach for studying the molecular evolution of pathogen host adaptation and pathogenesis, and facilitate

the tracing of its geographical spread (Figure 1.5) (Loman and Pallen, 2015, Bentley and Parkhill, 2015).

Early comparative genomic analysis of multiple pathogenic isolates of *Streptococcus agalactiae* revealed genome content variability among individual isolates, which then formalized the concept of pan-genome, describing the union of genes present in a species (Tettelin et al., 2005). Bacterial pan-genome is made up of core genes that are found in all members of a species and accessory genes that are strain-specific, found in only subset of strains (Tettelin et al., 2005). Evolutionary processes that contribute to the genome content variation of bacterial species, includes HGT, gene losses and gene duplications (McInerney et al., 2017). Pan-genome analyses provide a framework to determine genomic diversity among pathogen isolates associated with different clinical presentations, help to identify variations associated with pathogen virulence. For example, comparative analysis of clinical and non-clinical *Streptococcus suis* revealed that disease causing isolates had smaller genomes, with approximately 50 fewer genes than non-clinical isolates, and that genome reduction was associated with pathogenicity (Weinert et al., 2015). Furthermore, comparison of genomes from pathogen isolates from multiple host species can provide insights to the mechanisms and evolutionary process that drives host-adaptation. For example, a recent population-genomic analysis of over 800 *S. aureus* isolates from a diverse range of host species revealed that evolution of adaptation to specific hosts were contributed by various combinations of MGEs, presumably derived from the host-specific accessory gene pools through HGT (Richardson et al., 2018).

Genome analysis of bacterial species derived from different geographical locations over a period of time enables epidemiological investigations, which has the capacity to track patterns of spread and identify events which may have influenced the

spreading potential (Bentley and Parkhill, 2015). Such analysis provided evidence that contemporary *Burkholderia pseudomallei* isolates in Central and South American originated from Africa, and were potentially introduced to the Americas between 1650 and 1850, with the spread of the pathogen associated with the slave trade (Chewapreecha et al., 2017). Furthermore, epidemiological studies using WGS can facilitate endemic infectious disease outbreak investigation, to determine the source of outbreak and its local transmission. For example, using WGS and hospital surveillance data, an outbreak of MRSA in neonatal intensive units was traced to a healthcare worker as the carrier of the outbreak strain, who was likely to be responsible for the re-introduction of the strain to the unit following a deep clean (Harris et al., 2013).

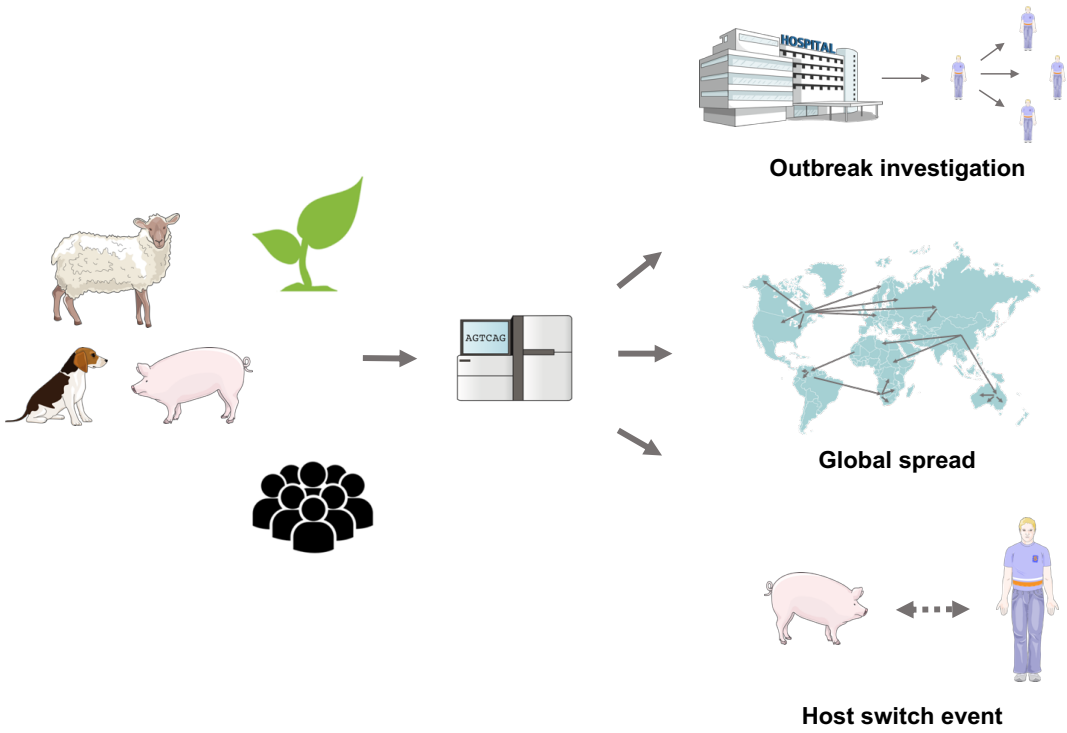


Figure 1.5 Bacterial genomics in infectious diseases. Sequencing of bacterial pathogens sampled from the environment, animals or human population can facilitate i) epidemiological investigation to identify the source of outbreak and subsequent local transmission event, ii) trace global spread of the pathogen and iii) evolution of pathogen virulence and host-adaptation, contributing host switching events. These analyses help to further our understanding of the evolution and spread of pathogen, in order to design intervention and therapeutics for control of the infectious disease.

1.4 Evolution and diversity of obligate intracellular bacteria

Unlike facultative intracellular bacteria, obligate intracellular bacteria have lost the capacity for replication outside their hosts and display unique host-microbe interactions (Toft and Andersson, 2010). During adaptation to the intracellular environment, various evolutionary processes have taken place which have shaped the bacterial genome and resulted in specialisation to a restricted niche.

1.4.1 Genome reduction in obligate intracellular bacteria

Genomes of obligate intracellular micro-organisms evolve under conditions that are very different to those of free-living organisms. First, the population dynamics within an intracellular environment is radically different to that of an extracellular environment, as growth is restricted by the limited cellular space, the number of infected cells and the number of hosts (Andersson and Kurland, 1998, Moran, 1996). Furthermore, population bottlenecks may be introduced during transmission between hosts (Moran, 1996). Each factor contributes to a reduction in the effective population size for obligate intracellular organisms, which will experience a greater effect of genetic drift due to relaxed purifying selection and a high fixation rate for deleterious mutations (Moran, 1996, Funk et al., 2001). Consequently, this leads to accelerated sequence evolution and increased divergence from the ancestral population. Second, the intracellular environment is metabolite rich and adaptation to the host therefore leads to relaxed selective constraints, extensive pseudogenisation and gradual loss of genes involved in biosynthesis of products that can be obtained from the host (Moran, 1996, Moran et al., 2009). Third, the isolated intracellular environment restricts the opportunity for gene acquisition as access to novel gene pools is

constrained, and HGT occurs less frequently than compared to free-living bacteria (Bordenstein and Reznikoff, 2005).

Gradual gene loss combined with the restricted influx of novel genes leads to the phenomenon of genome reduction over time, in which the extent of irreversible gene loss accompanies host dependency (Toft and Andersson, 2010). Consequently, the different stages of bacterial adaptation into the host intracellular environment will correlate with a reduction in its genome size, a common trend among all clades of intracellular bacteria (Toft and Andersson, 2010). Such phenomenon has been demonstrated in *L. intracellularis* through sequencing of the N343 strain, which revealed components of the tricarboxylic acid cycle and genes involved in synthesis of specific amino acids to be absent (Sait et al., 2013), most likely as the result of genome decay.

1.4.2 Mutualistic versus parasitic association

During the course of adaptation to the intracellular environment, the bacterium may form either a mutualistic or parasitic relationship with their host (Moya et al., 2008). A mutualistic relationship is formed when the bacterium and its host become dependent on one another for survival (Toft and Andersson, 2010). This is caused by gradual loss of gene functions from both partners as a consequence of overlapping functional redundancy, leading to an endosymbiotic relationship (Moya et al., 2008). Endosymbionts are often transmitted through maternal inheritance (vertical transmission) which facilitates a more stable and effective transmission route among host generations (Bright and Bulgheresi, 2010, Toft and Andersson, 2010). On the other hand, symbionts with a parasitic association with their host are often transmitted

from one individual to another (horizontal transmission) (Bordenstein and Reznikoff, 2005). A parasitic relationship will result in decreased fitness of the host in which damage may be caused by the microbe, the host immune response or both. During co-evolution of pathogen and hosts, adaptation may result in the gradual loss of virulence if survival of the pathogen is dependent on the survival of the host (Casadevall, 2008). However, if host damage and death is required for transmission, a co-evolutionary arms race between the bacterium and host immune system may eventually lead to selection for genes involved in immune evasion and enhanced intracellular survival of pathogen (Casadevall, 2008, Casadevall and Pirofski, 1999).

1.4.3 Horizontal gene transfer in obligate intracellular pathogens

It is commonly assumed that HGTs are unlikely to occur in obligate intracellular bacteria due to their isolated intracellular niche restricting exposure to novel gene pools, representing a barrier for recombination. However, with the availability of WGS the idea that obligate intracellular bacteria are impervious to HGT is being challenged. Comparative whole genome analysis of a global collection of *Chlamydia trachomatis* isolates revealed frequent homologous recombination occurring within and between strains from the two biovars, which have tropisms for different tissues (Harris et al., 2012). Furthermore, mobile genetic elements have been previously identified in *Wolbachia*, comprising up to 21% of its genome, with evidence of ongoing HGT events between different strains facilitated by bacteriophage WO during coinfections (Chafee et al., 2009, Bordenstein and Wernegreen, 2004). This supports the 'intracellular arena' hypothesis, by which genetic material can be exchanged between communities of obligate intracellular bacteria co-infecting the same intracellular host environment (Bordenstein and Wernegreen, 2004).

1.5 Introduction to *Lawsonia intracellularis*

The Gram-negative, obligate intracellular bacterium *Lawsonia intracellularis* (*L. intracellularis*) is the aetiological agent of a non-zoonotic enteric disease known as proliferative enteropathy (PE) (Lawson and Gebhart, 2000). The disease was first described in pigs in 1931 (Biester and Schwarte, 1931), but it was not until the early 1970s that the presence of this intracellular bacteria was identified in proliferative lesions of affected pigs (Lawson and Rowland, 1974).

Initially, *L. intracellularis* were referred to as '*Campylobacter*-like organisms' due to morphological similarities with *Campylobacter* species (McOrist & Gebhart, 1999). However, experimental infection with several *Campylobacter* species recovered from PE lesions failed to reproduce the disease. Furthermore, anti-sera derived from rabbits immunised with bacterial extracts isolated from PE lesions was unreactive against *Campylobacter spp*, leading to the speculation that a novel intracellular bacterium was the causative agent of PE (Lawson et al., 1985). Following success in isolation and cultivation of this intracellular bacteria, G.H.K Lawsons' group was able to reproduce the disease, fulfilling Koch's postulates (Lawson et al., 1993a, McOrist et al., 1993a). The bacterium was first given the name Ileal Symbiont (IS) intracellularis and later *Lawsonia intracellularis* in honour of its primary discoverer, Dr. Lawson (McOrist et al. 1995).

1.5.1 Clinical characteristics of proliferative enteropathy

All cases of PE are characterised by a thickening of the intestinal mucosa and hyperplastic lesions due to proliferation of the enterocytes. However, clinical and pathological presentations may vary among different host species. In pigs, two clinical

manifestations have been described; i) a milder, self-limiting form known as porcine intestinal adenomatosis (PIA) commonly affecting weaners or young growing animals between 6 and 20 weeks of age, and ii) an acute form known as proliferative haemorrhagic enteropathy (PHE) more commonly observed in mature animals between 4 to 12 months of age (McOrist and Gebhart, 2012). The mortality of animals suffering from PHE is high and may reach 50% (Love et al., 1977), predominant clinical signs are bloody diarrhoea and intestinal haemorrhage (McOrist and Gebhart, 2012). Furthermore, sudden death of animals without faecal abnormality has often been reported (Love et al., 1977). In contrast, the mortality of PIA is usually low, around 1-5% (Jacobson et al., 2010). Clinical signs of PIA include diarrhea, proliferation of mucosal cells in the intestine and decreased weight gain (McOrist and Gebhart, 2012). Most cases of PIA can be sub-clinical by which animals do not exhibit any clinical presentations, but may result in decrease of average daily weight gain (Paradis et al., 2012).

Equine PE commonly affects foals aged between 2 - 8 months and occasionally adult animals. Clinical presentation includes peripheral oedema, anorexia, fever, colic, diarrhoea and lethargy (Lavoie et al., 2000). Similar to PIA, the disease can often be sub-clinical but animals may continue to decrease in average weight gain (Lavoie et al., 2000). A consistent clinical presentation of equine PE is hypo-proteinemia due to hypalbuminaemia, which has not been described in pigs. In some cases, hypo-proteinemia may be the only clinicopathological abnormality presented (Frazer, 2008).

1.5.2 Pathogenesis of *L. intracellularis* infection

During early stage of pathogenesis, *L. intracellularis* targets immature enterocytes at the intestinal crypts, although infection in mature enterocytes can sometimes be observed (Boutrup et al., 2010). Although factors required for *L. intracellularis* adherence and entry into host cells remains undetermined, *in vitro* infection studies have previously demonstrated that the invasion process involves actin polymerisation and rearrangement, independent of bacterium viability (Lawson et al., 1995). Upon internalisation, the bacterium is initially associated with membrane-bound vacuoles but quickly escapes, by which the bacterium can be found lying free in the cytoplasm 3 hours after exposure (McOrist et al., 1995). Within the intracellular environment, *L. intracellularis* is typically found residing in the apical cytoplasm and replication through binary fission can be observed two to six days after infection *in vitro* (McOrist et al., 1995).

Transcriptional profiling of *L. intracellularis* from *in vivo* and *in vitro* infection, has demonstrated high levels of expression of genes including redox enzymes, Cu-Zn superoxide dismutase (*sodC*) and rubrerythrin-rubredoxin operon (*rubY* - *rubA*) (Vannucci et al., 2012a, Vannucci et al., 2013c). While redox enzymes catalyse the reduction of O_2, *sodC* and *rubY* – *rubA* can neutralize reactive oxygen species (Lumppio et al., 2001), and it is thought these mechanisms of coping with oxidative stress may facilitate the intracellular survival of *L. intracellularis* (Vannucci et al., 2012a, Vannucci et al., 2013c). Furthermore, a gene encoding a putative ATP/ADP translocase, involved in catalysing the exchange of bacterial ADP for host ATP in two obligate intracellular bacteria (*Chlamydiae* and *Rickettsiae*), have been identified in *L. intracellularis*, suggesting the organism may utilise the host energy pool in a similar manner (Schmitz-Esser et al., 2008).

During active *L. intracellularis* infection, infected cells become proliferative and hyperplasic. Evaluation of the host transcriptional response has demonstrated up-regulation of genes involved in cell cycle regulation and cell differentiation, including Rho family genes and CDK1 which drives cell cycle progression into G_1 and G_2 phase, respectively (Smith et al., 2014, Vannucci et al., 2013b). This active proliferation of infected cells in turn facilitates the rapid spread of the bacterium throughout the epithelium, as *L. intracellularis* replication occur concurrently with enterocyte division (McOrist et al., 1995). At the same time, significant down-regulation of genes for nutrient acquisition has been observed during *in vivo* and *in vitro* infection, including genes encoding sodium/glucose transporters, cationic amino acid transporters, ileal sodium/bile acid transporters, lipid phosphate phosphohydrolases and various members of the solute carrier family (Vannucci et al., 2013c, Smith et al., 2014). Furthermore, transcripts for the S100 calcium binding protein G (calbindin d9K), which facilitates intestinal absorption and transport of calcium, demonstrated the greatest fold down-regulation in ileal tissues of pigs experimentally infected with *L. intracellularis* (Smith et al., 2014). The repressed expression of genes for nutrient acquisition leads to malabsorption of the infected cell and prevents enterocyte differentiation into mature cells (Smith et al., 2014, Vannucci et al., 2013b, Oh et al., 2010). Consequently the intestinal mucosa is crowded with proliferating immature epithelial cells (McOrist et al., 1996). These alterations in host gene expression upon *L. intracellularis* infection ultimately lead to poor performance and growth of the animal.

Examination of signalling pathways during *L. intracellularis* infection *in vivo* demonstrated a simultaneous induction of Notch-1 signalling and down-regulation of the β-catenin/Wnt pathway, preventing goblet cell maturation and enhanced crypt cell proliferation (Huan et al., 2017). This coincides with the depletion of goblet cells and

suppressed crypt epithelial cell maturation during the peak of infection (Bengtsson et al., 2015, Smith et al., 2014). Furthermore, a down-regulation of genes involved in intestinal mucosal integrity is observed, diminishing the host mucosal integrity and may provide an opportunity for *L. intracellularis* to further invade epithelial cells (Bengtsson et al., 2015, Smith et al., 2014).

The development of PHE, the haemorrhagic form of the disease, has been associated with high levels of macrophages carrying *L. intracellularis* present within the lamina propria, submucosa, epithelial capillaries and lymphatics (Love and Love, 1979). Hence, it has been suggested that macrophages may play a role in facilitating the spread of the organism and inducing an acute inflammatory response, absent in cases of PIA (Love and Love, 1979). During the healing stages of PE, the lack of *L. intracellularis* observed in epithelial cells is thought to be associated with the resumption of apoptotic events and the restoration of normal intestinal mucosal structure (McOrist et al., 1996).

1.5.3 Transmission of *L. intracellularis*

The primary mode of disease transmission of *L. intracellularis* is by the faecal-oral route (Guedes, 2004). *L. intracellularis* can persist within an animal host over long period up to 12 weeks, with irregular shedding of the organism from infected animals (Guedes and Gebhart, 2003a). Studies on the survival of *L. intracellularis* outside the host revealed that the microbe remained viable for up to 14 days in porcine faeces between 5-15^{o}C (Collins et al., 2000). In addition, the disease can be maintained by sub-clinically infected pigs (Jacobson et al., 2003). The long-term persistence of *L. intracellularis* in the environment and within the host permits reinfection among swine herds.

Evidence of exposure to *L. intracellularis* has been demonstrated in rats and mice captured from farms with PPE and EPE cases (Hwang et al., 2017, Pusterla et al., 2008, Friedman et al., 2008). While PCR assessment of *L. intracellularis* in faeces of wild rodents from farms with EPE cases, demonstrated low prevalence of 3% (Hwang et al., 2017). Study examining the presence of *L. intracellularis* DNA in faeces of wild rats trapped in pig farms with endemic PE, demonstrated high prevalence of more than 76.6% PCR-positive animals in which a small proportion were observed to shed up to 10^{10} *L. intracellularis* per gram of faeces (Collins et al., 2011), an sufficient amount to cause infection in pigs (Collins and Love, 2007). In addition, detection of *L. intracellularis* DNA in faeces of stray cats and rabbits on farms with confirmed EPE has been previously reported (Hwang et al., 2017, Pusterla et al., 2012). These findings suggests that wild animals may act as biological reservoirs for *L. intracellularis* and may play a role in the introduction, reintroduction and maintenance of the organism within farms (Friedman et al., 2008, Pearson et al., 2016).

1.5.4 Host association of *L. intracellularis*

A cross-species experimental model using pigs and horses infected with pure culture of either a porcine or equine-derived *L. intracellularis* strain demonstrated host specificity of infection (Figure 1.6) (Sampieri et al., 2013b, Vannucci et al., 2012b). While pigs challenged with a porcine-derived *L. intracellularis* strain developed macroscopic and histologic lesions typical of porcine PE, animals challenged with the equine-derived strain displayed no clinical signs or pathological changes (Vannucci et al., 2012b). Furthermore, pigs infected with the porcine-derived strain demonstrated a stronger serological immune response for *L. intracellularis,* and a higher and longer period of bacterial shedding compared to pigs infected with the equine derived strain (Vannucci et al., 2012b). The authors observed the same species-specific infection outcome in the challenge model in foals (Vannucci et al., 2012c).

Although exposure to *L. intracellularis* has been identified in rodents from farms with EPE cases, animals captured appeared clinically normal with no signs of adenomatous lesions (Hwang et al., 2017). However, intestinal mucosal tissues of a small proportion of animals were PCR-positive for *L. intracellularis*, suggesting potential transmission infection between rodents and horses (Hwang et al., 2017). A recent study investigating the faecal-oral transmission of *L. intracellularis* between mice and pigs demonstrated similar findings, in which mice exposed to porcine faeces were actively shedding the bacterium but did not display clinical signs and gross lesions (Gabardo et al., 2017). In contrast, pigs exposed to faeces of *L. intracellularis* positive mice developed active infection displaying poor clinical conditions and macroscopic lesions in which presence of *L. intracellularis* were confirmed by IHC (Gabardo et al., 2017).

Hamsters are known to be naturally infected by *L. intracellularis*, displaying typically clinical presentation of PE including diarrhoea, weight loss and severe hyperplasia in the small intestine (Stills, 1991). Hence, hamsters are often used as a model for porcine PE infection studies, by which animals challenged with a pure culture of porcine derived *L. intracellularis* strain consistently reproduced PE (Jasni et al., 1994, Vannucci et al., 2010, Sampieri et al., 2013b). In contrast, hamsters challenged with a pure culture of equine-derived *L. intracellularis* did not develop infection (Sampieri et al., 2013b). Experimental infection using rabbits demonstrated the opposite effect, by which the animal developed intestinal lesions, an immune response and faecal shedding when challenged with an equine strain, but appeared clinically normal when challenged with a porcine strain (Figure 1.6) (Sampieri et al., 2013a, Sampieri et al., 2013b).

Early studies investigating 16S ribosomal DNA sequences of *L. intracellularis* derived from pigs, hamsters, deers, ostriches and ferrets, revealed that isolates all shared high sequence similarity and are all *L. intracellularis* (Cooper et al., 1997b, Fox et al., 1994). Taken together, these observations suggest that *L. intracellularis* may exist as variants that have evolved to infect more than one host species. However, the genetic basis responsible for the host-specific characteristic of *L. intracellularis* remains undetermined.

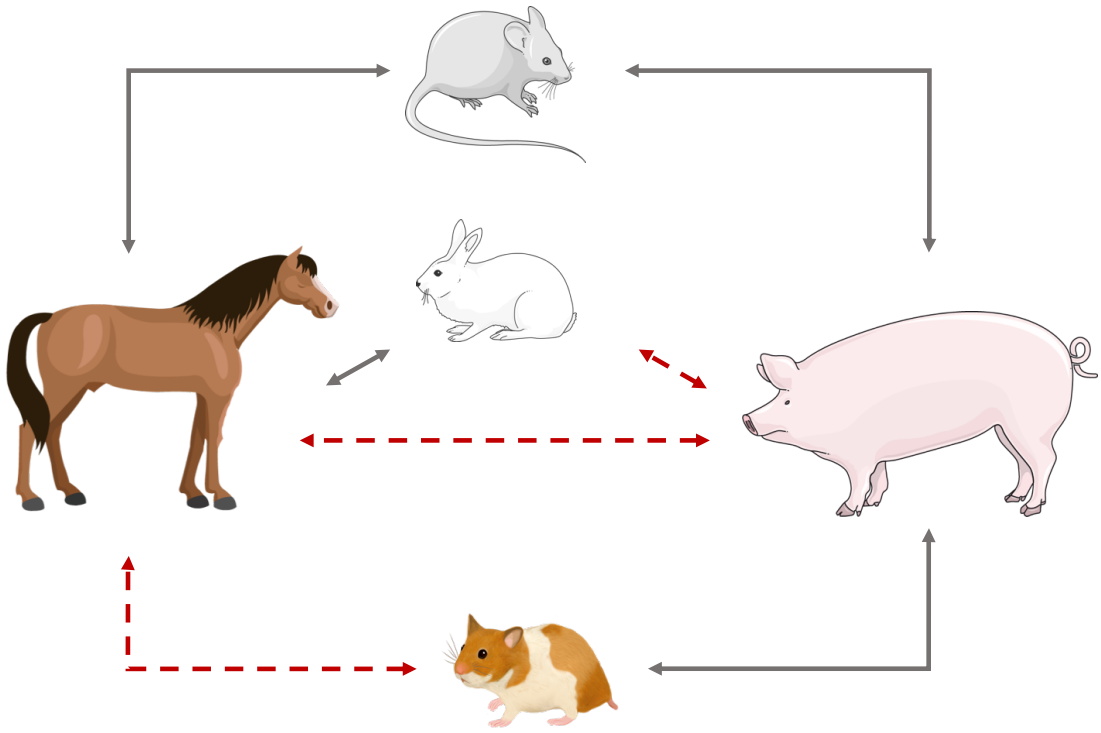


Figure 1.6 Schematic diagram displaying cross-species infection of *L. intracellularis*. Cross-species *L. intracellularis* experimental infection with equine derived strain E40504 and porcine derived strain PHE/MN1-00, in different animal models demonstrated host specificity for infection. Red dotted lines represent inability of isolate derived from either equine or porcine host to cause infection in another host. Figure constructed based on data from Vannucci et al., 2012b and Sampieri et al., 2013b.

1.5.5 Genomic features of *L. intracellularis*

L. intracellularis belongs to the Desulfovibrionaceae family of the class deltaproteobacteria, and is the sole species described, to date, for the genus *Lawsonia* (Gebhart et al., 1993). Previously, the sulphate-reducing bacterium *Desulfovibrio desufuricans* was thought to be the closest relative to *L. intracellularis*, based on 16S ribosomal DNA sequence (Gebhart et al., 1993). However, a recent large scale phylogenomic analysis based on alignment of 15 ribosomal proteins, that included over 1,000 uncultivated organisms together with published genomes, demonstrated that *L. intracellularis* was most closely related to *Bilophila wadsworthia* (Hug et al., 2016), a Gram-negative commensal gut bacterium thought to cause appendicitis of humans (Baron, 1997).

To date, six *L. intracellularis* genome sequences are available on NCBI, PHE/MN1-00, N343, E40504, Ni/JPN, Fu/JPN and Ib2/JPN. Strain PHE/MN1-00 was the first complete genome sequenced using Sanger-based sequencing in 2006 followed by the complete genome sequence of strain N343 in 2013 (Sait et al., 2013). Both PHE/MN1-00 and N343 strain were isolated from pigs with haemorrhagic PE, whole genome comparative analysis demonstrated limited genetic variation between the two strains with 24 SNPs and 90 indels (Sait et al., 2013). Three Japanese strains Ni/JPN, Fu/JPN and Ib2/JPN were recently deposited into the public database. A draft genome of an equine strain, E40504, has also been publicly available since 2017 (Mirajkar et al., 2017). A comparison of equine and porcine origin strains showed 99.63% nucleotide identity (Mirajkar et al., 2017). All sequenced *L. intracellularis* genomes comprise a 1.4 Mb chromosome and 3 plasmids of 27, 39 and 194 Kbp, respectively, with a low GC content of 33%. The genome encodes two rRNA gene copies and approximately 1,340 protein-coding genes (Sait et al., 2013, Mirajkar et

al., 2017). Further analysis of *L. intracellularis* has identified genes predicted to be involved in virulence, adhesion and invasion, including type III secretion system (TTSS) proteins, type V autotransporter (AT) proteins and flagellum assembly proteins (Sait et al., 2013). Due to difficulties associated with growth of the bacterium very few genes/proteins have been experimentally characterised.

Currently, there is limited understanding of the molecular mechanisms for *L. intracellularis* infection and virulence. Attenuation of virulence in the Sanger-sequenced *L. intracellularis* PHE/MN1-00 strain was achieved at passage number 40, by which pigs challenged with pure culture of isolates at 40 passages did not develop clinical signs of PE (Vannucci et al., 2013a). A lack of genetic variation was observed when comparative genomic analysis of the same strain cultivated for 10 and 60 passages demonstrated only five SNPs differences across the entire genome, and a deletion of an 18 kb prophage- associated genomic island in the passage 60 isolate (Vannucci et al., 2013d). A study examining transcriptional profiling of the 10 and 60-times passaged PHE/MN1-00 isolates revealed alteration in the attenuated variant at passage 60, by which expression of genes involved in membrane transport, adaptation and stress response were repressed (Vannucci et al., 2012a). Hence, it is speculated that mechanisms to attenuate virulence properties of *L. intracellularis* in the high passaged isolate may occur through epigenetic regulation which leads to gene silencing events (Vannucci et al., 2012a). Furthermore, transcriptional profiling of *L. intracellularis* during *in vitro* and *in vivo* infection identified high level of expression in genes encoding hypothetical proteins, indicating that this organism may possess unique mechanisms for survival and pathogenesis (Vannucci et al., 2012a, Vannucci et al., 2013b).

1.5.6 Clinical and economical significance of *L. intracellularis*

Porcine PE is an endemic disease with a high herd prevalence in all major pig producing countries. In Denmark, 93.7% of herds tested positive by PCR for *L. intracellularis* (Stege et al., 2000) and 48% in piglet producing herds in Sweden (Jacobson et al., 2005). In South Korea and Australia, 100% and 84.2% of herds tested positive for serum antibodies against *L. intracellularis*, respectively (Lee et al., 2001, Holyoake et al., 2010). Among intensive pig farms in China the prevalence of *L. intracellularis* was at 77% seropositivity (Wu et al., 2014). The disease has a significant impact on health, welfare and animal production, imposing a huge economic burden. For chronic cases, feeding conversion efficiency of animals may be reduced by up to 50% with total losses estimated to be approximately €5 per affected animal (Gogolewski et al., 1991, Schnurrbush, 2014), and loss of €11 per affected animal has been estimated for acute cases (Schnurrbush, 2014).

1.5.7 Control of *L. intracellularis*

Currently, commercial vaccines against *L. intracellularis* are available for prophylactic use, this includes the live attenuated vaccine Enterisol® Ileitis and an inactivated bacterin-based vaccine (Roerink et al., 2018, McOrist and Smits, 2007). Although numerous studies have demonstrated the efficacy of Enterisol® Ileitis, with vaccinated animals demonstrating reduced bacterial shedding and intestinal lesions upon re-challenge (Guedes and Gebhart, 2003b, Kroll et al., 2004, Riber et al., 2015, Park et al., 2013, Peiponen et al., 2018). Vaccination with Enterisol® Ileitis does not confer sterile immunity, whereas animals with natural infection are protected from re-colonisation and resistance to reinfection (Riber et al., 2011, Collins and Love, 2007).

Furthermore, vaccination with Enterisol® Ileitis did not prevent *L. intracellularis* faecal shedding in re-challenged animals (Guedes and Gebhart, 2003b, Kroll et al., 2004). Finally, development of protective immunity is observed three to four weeks following administration of Enterisol® Ileitis (Walter et al., 2004), and during this period antibiotic use is prohibited due to the live nature of the vaccine. More recently, an inactivated vaccine, Porcilis® Ileitis, has recently been licenced for commercial use. In vaccine trials of farrow-to-finish pigs from a commercial farm, the vaccine provided protection against experimental challenge with *L. intracellularis* (Roerink et al., 2018). However, due to only recent availability of Porcilis® Ileitis, there is no information yet in the literature regarding the immunity induced and duration of protection provided by the vaccine.

1.6 Background summary

Proliferative enteropathy caused by the obligate intracellular pathogen *L. intracellularis* have been identified in both wild and domestic animals. The disease has significant impact on health and welfare of animals, imposing huge economic burden on the pig farming industry. Currently, only six genome sequences are available for *L. intracellularis.* The mechanisms for inducing proliferative changes during infection, and the genetic determinants for pathogenesis and host specificity for infection remains largely undetermined.

Hight throughput sequencing of bacterial genomes has significantly increased our understanding on the biology of many important pathogens, endangering public health and animal welfare. Population genomics of bacterial pathogens can provide insights to the evolutionary forces and processes that drives emergence of epidemic clones, identify genetic determinants for pathogenesis and facilitate epidemiological investigation to trace outbreak origin and transmission across community. The wealth of knowledge generated from genomic data are important to inform development of therapeutic treatments and design intervention to control infectious diseases. Advances in HTS technologies, together with new development of computational approaches has accelerated the field of metagenomics and contributed to the expansion of near-complete microbial genomes deposited in public databases. Despite methodological and practical challenges, studies have demonstrated the potential use of metagenomics in clinical microbiology for pathogen detection and discovery.

1.7 Project Aims

The aim of this project was to utilise genomics in order to characterise *L. intracellularis*, to help inform the development of therapeutic treatments for the control of PE.

1. Obtain draft genome sequences of *L. intracellularis* through direct sequencing of clinical samples derived from pigs and horses with PE.

2. Investigate the genetic diversity of *L. intracellularis* isolates, identify genetic determinants for pathogenesis and host specificity.

3. Perform immunological investigation of a novel surface protein.

Chapter 2

Recovery of *Lawsonia intracellularis* draft genomes by direct sequencing of clinical sample

2.1 Introduction

Due to its obligate intracellular and microaerophilic lifestyle, *L. intracellularis* has extremely fastidious *in vitro* growth requirements, which makes bacterium culture labour intensive and challenging (Vannucci et al., 2012d). Hence, generating sufficient DNA for traditional WGS is difficult for this bacterium. For this reason, there is a lack of understanding on the genome biology of this pathogen. To date, six *L. intracellularis* genome sequences are available on NCBI, all obtained from sequencing of cell cultured samples.

Metagenomic sequencing bypasses the need to culture prior to sequencing, and has been previously utilised to recover genomes of several intracellular pathogens, including chlamydial species from clinical swabs of snakes, *Mycobacterium leprae* from human skin biopsies and *Mycobacterium tuberculosis* from sputum samples (Taylor-Brown et al., 2017, Benjak et al., 2018, Doughty et al., 2014). However, analysis is challenging owing to the presence of repeated sequences within and across genomes, highly non-uniform coverage, sequencing errors and strain-level variation within the microbial community (Sczyrba et al., 2017, Vollmers et al., 2017). Numerous metagenomic-specific assembly and sequence analysis software been developed to address the various challenges of metagenomic data. However, there is not a single tool that will satisfy all current shortcomings, as each comes with its own limitations.

In the current study, we explored metagenomic sequencing of faecal and ileum tissue samples derived from pigs and horses with PE, in order to obtain *L. intracellularis* genomes. In addition, we assessed the performance of various tools and methods in

order to develop a bioinformatic pipeline to acquire high-quality draft genomes of *L. intracellularis* from metagenomic data.

2.2 Aims

1. Acquire *L. intracellularis* sequences through direct sequencing of clinical samples.

2. Enrich microbial DNA from clinical sample to increase *L. intracellularis* sequencing depth of coverage.

3. Develop bioinformatic pipeline to assemble and recover high quality draft genomes from metagenomic samples by:

    a. Assessing performance of metagenomic assemblers on *L. intracellularis* genome assembly

    b. Comparing supervised and unsupervised binning to maximise *L. intracellularis* genome recovery

2.3 Materials and Methods

2.3.1 Bacterial strains, DNA extraction and microbial DNA enrichment

All samples were stored in -80°C freezer prior to DNA extraction. Faecal DNA extraction was performed using DNeasy PowerSoil kit (Qiagen) according to manufacture instructions. Samples placed in PowerBead tubes were placed in FastPrep® homogenizer (MP Biomedicals) for 1 min to homogenise sample and mechanically lyse bacterial cells. DNA from Ileum and colon samples was extracted using DNeasy Blood and Tissue Kits (Qiagen) following the purification of total DNA from animal tissues (spin-column) using the protocol provided in the manufacturer's instruction handbook. DNA from cell-cultured samples was extracted using DNeasy Blood and Tissue Kits (Qiagen) according to the protocol for purification of total DNA from animal blood or cells (spin-column). Enrichment of microbial DNA from genomic samples was carried out using the NEBNext® Microbiome DNA Enrichment Kit (New England Biolab).

The quality and integrity of DNA was measured using the Aligent 4200 TapeStation System (Agilent Genomics) and NanoDrop™ 1000 spectrophotometer (Thermo Fisher Scientific), and the quantity of DNA was measured using Qubit® 3.0 fluorometer with Qubit dsDNA BR Assay Kit (Invitrogen).

2.3.2 Quantitative Polymerase Chain Reaction

L. intracellularis genomic DNA from clinical samples was quantified using 10 μl of LightCycler®480 SYBR Green I Master Mix (Roche) 10 μM of forward and reverse primers, 7.2 μl of water, PCR-grade (Roche) and 2 μl of template DNA, to a final

volume of 20 μl. The thermocycler programme used on a LightCycler®480 instrument II (Roche) included an initial pre-incubation at 95°C for 5 min, followed by amplification of target DNA with 45 cycles of denaturation at 95°C for 10 s, annealing at 52°C for 13 s, extension at 72°C for 9 s, followed by melting curve for PCR product identification at 95°C for 5 s and 65°C at 65°C for 1 min. Finally, the multi-well plate was cooled at 40°C for 30 s. *L. intracellularis* genomic DNA was quantified by primers specific for the *aspA* gene (Invitrogen) and samples were tested in triplicates. Tenfold serial dilutions of gBlocks®Gene fragment comprising of *aspA* sequence (150bp) were used to construct a standard curve to quantify *aspA* copies in qPCR assay. Primer and gBlocks®Gene fragment sequences are provided in Table 2.1

Quantification of *L. intracellularis* genomic DNA from cell cultured samples was performed using the same LightCycler®480 SYBR Green I Master Mix kit and followed the same thermocycler programme and procedure as previously described, except for an annealing temperature of 60°C for 13 s and extension at 72°C for 13 s. *L. intracellularis* genomic DNA was detected using primers targeting the 16S rRNA gene (Invitrogen) and quantified using a standard curve plasmid, obtained from a previous study by (Smith et al., 2014), in which plasmid was constructed using a pGEM®-T vector plasmid containing a 322 bp *L. intracellularis* ribosomal 16S rRNA gene insert (Smith et al., 2014).

2.3.3 Genome sequencing and sequence processing

Genomic DNA libraries were prepared using Illumina TruSeq Nano 550bp Gel Free kit (Illumina). Library preparation and sequencing service was provided by Edinburgh Genomics. Sequencing was performed on the Illumina Hi-Seq 2500, Hi-Seq 4000 or

Mi-Seq instruments (Table 2.2). Samples were multiplexed, ranging from 6 to 4 samples per lane.

Quality control checks on raw sequence data were carried out using FastQC (Andrews, 2010). All raw reads were adapter- and quality-trimmed with Trimmomatic v0.36 on paired-end mode (Bolger et al., 2014).

2.3.4 Taxonomic classification and host reads removal

Following adapter trimming and quality filtering of raw reads, Kraken v2.0 was used to remove contaminating host reads from filtered sequence data. A custom kraken database containing host reference genomes (krakenHost database) was constructed, comprising of domestic pig (*Sus scrofa*, GenBank GCA_ 000003025.6), rat (*Rattus norvegicus*, GenBank GCA_000001895.4), mouse (*Mus musculus*, GenBank GCA_000001635.8) and horse (*Equus caballus*, GenBank GCA_002863925.1) genomes (Wood and Salzberg, 2014). Filtered reads were mapped against the krakenHost database, reads classified by Kraken as host were discarded while unclassified reads were used for downstream analyses.

For taxonomic classification of the unclassified reads, sequence data were mapped against the Standard kraken 2 database, containing complete genomes in RefSeq for the bacterial, archaeal and viral domains (Wood and Salzberg, 2014).

2.3.5 Genome assembly and annotation

For metagenomic assembly, the processed sequences were used as input for assembly using MEGAHIT, metaSPAde and IDBA-UD (Li et al., 2015, Nurk et al., 2017, Peng et al., 2012).

Unsupervised binning of the assembled contigs was carried out using MetaBAT2 (Kang et al., 2015). Prior to binning, BWA MEM was used to map processed sequence reads onto assembled contigs (Li, 2013). Contig coverage from each assembly was calculated using the script jgi_summarize_bam_contig_depths from the MetaBAT2 package, followed by binning of contigs using option --minContigLength 2000, --minContigDepth 2. CheckM was used to assess genome bin quality based on single copy lineage specific marker genes (Parks et al., 2015).

Supervised binning of the assembled contigs was carried out using QUAST v4.6.0 using the reference genome sequence of PHE/MN1-00 (GenBank GCA_000055945.1). Aligned contig headers were retrieved from the contigs.tsv file in the contigs_reports folder produced by QUAST. Using an in-house python script, genome sequences of the reference aligned contigs headers were extracted from the assembly output containing MAGs.

For the reference-guided assembly, the processed sequences were mapped to a *L. intracellularis* custom kraken database (krakenLI database) comprising of E40504 (GenBank GCA_001975945.1), N343 (GenBank GCA_000331715.1) and PHE/MN1-00 (GenBank GCA_000055945.1) genomes retrieved from NCBI. The *L. intracellularis* classified reads mapped to the krakenLI database were used as input

for metaSPAdes genome assembler v3.11.1 (Bankevich et al., 2012). Genome sequences were annotated using Prokka (Seemann, 2014).

2.3.6 Genome assembly quality assessments

The quality and statistics of genome assembly were evaluated with QUAST v4.6.0 using the complete genome sequence of PHE/MN1-00 as reference for porcine isolates and the draft genome of E40504 as reference for equine isolates (Gurevich et al., 2013).

Table 2.1 Primers and gBlock gene fragment sequence employed in the current study.

Primer Name	Sequence (5' to 3')	Gene target
16SF	GCGCGCGTAGGTGGTTA	16S rRNA
16SR	GCCACCCTCTCCGATACTCA	16S rRNA
LIC058F	TGTGTATATTTGACAGCTGGAGA	LIC058 (*LI_RS07080*)
LIC058R	CAAGGACGACGGCTTATCC	LIC058 (*LI_RS07080*)
*aspA*1F	AGATACGGGTGCTTATGTTCAG	*aspA*
*aspA*1R	GGTGCCCTTGGAGGTAAAT	*aspA*
aspA gBlock®Gene fragment	CACAAGATACGGGTGCTTATGTTCAGCTTTCTGGTGTTCTTA AAAGAGTTACAGTTAAACTTTCTAAAATCTGTAATGACTTACG ACTACTCTCAAGTGGTCCTCGCTGTGGATTGGGAGAAATCAA TTTACCTCCAAGGGCACCTGGTT	

2.4 Results

2.4.1 Development of a quantitative assay to measure *L. intracellularis* DNA from complex samples

To ensure sufficient quantity of *L. intracellularis* DNA is present prior to sample sequencing, a quantitative assay using qPCR targeting a *L. intracellularis* specific gene was developed. The 16S rRNA gene is ubiquitous among bacteria and its sequence is comprised of conserved and hypervariable regions (V1-V9), by which the variable regions are species-specific and is often used as a genetic marker for taxonomic classification of bacteria. Diagnostic tests for proliferative enteropathy often involve PCR assays to the detect presence of the bacterium in faecal samples, based on the 16S rRNA gene (La et al., 2006). PCR detection of *L. intracellularis* from cell cultured and ileum samples using 16S rRNA primers obtained from the study by Sionagh *et al* (Smith et al., 2014), resulted in amplification of an expected 250 bp product (Figure 2.1). Initially, these primers were employed to quantify *L. intracellularis* copy numbers within each sample, of which samples with the highest bacterium load were selected for sequencing. However, *L. intracellularis* copy numbers quantified using 16s rRNA gene did not correlate to the bacterium sequencing depth of coverage (Table 2.2). Subsequent PCR analysis with faecal DNA samples revealed non-specific amplification of multiple products, suggesting when used in samples containing rich microbial diversity the primer pair may anneal non-specifically to 16S rRNA gene of other bacteria within the sample, producing PCR products of various length (Figure 2.1).

Instead, primers targeting the *aspA* (*LI_RS04975*) gene, a single copy house-keeping gene located on the main chromosome of *L. intracellularis* which encodes for

aspartate ammonia-lyase (Wattanaphansak et al., 2010, Pusterla et al., 2010), were evaluated for specificity in detecting *L. intracellularis* from faecal DNA. The optimised PCR and qPCR assay targeting this gene resulted in amplification of a single 141 bp product of excepted size from faecal samples, suggesting the *aspA* primer pair are specific for *L. intracellularis* DNA in complex samples.

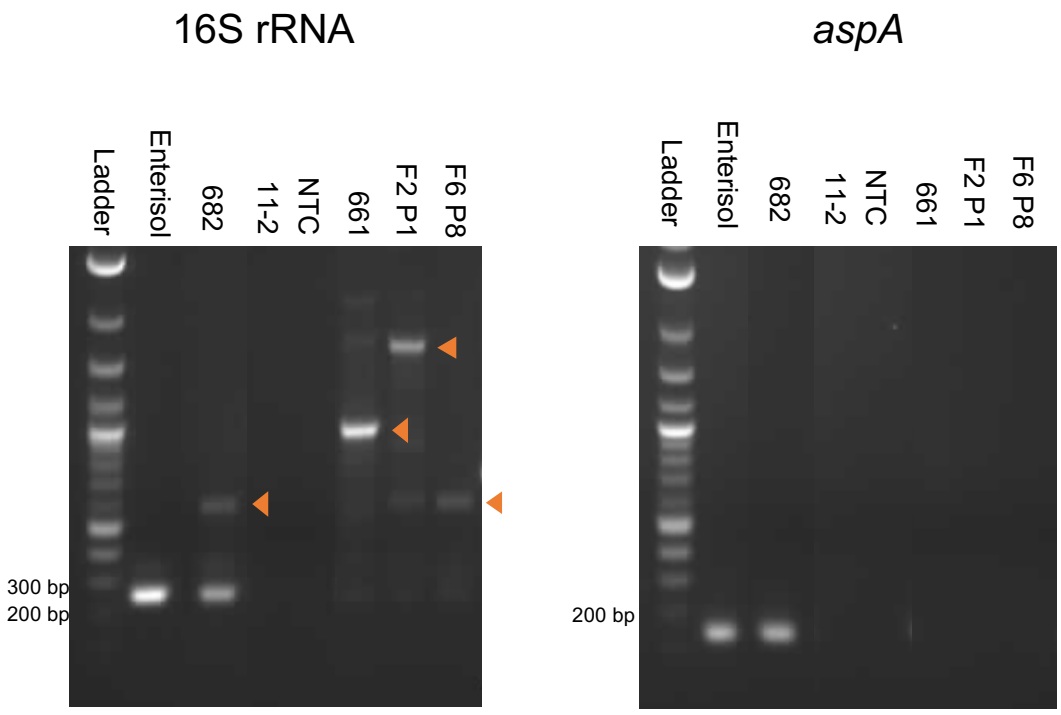


Figure 2.1 PCR amplification of *L. intracellularis* 16S rRNA and *aspA*. Agarose gel (2%) image of PCR amplified products using primer pairs targeting *L. intracellularis* 16S rRNA and *aspA* from DNA extractions. Lanes are arranged as follow; Enterisol® Ileitis is a cell-cultured vaccine strain, 682 is a faecal DNA sample from a positive for *L. intracellularis* infection, 11-2 is a faecal DNA sample from a pig negative for *L. intracellularis*, NTC (non-template control), 661, F2P1 and F6P8 are faecal DNA samples from pigs negative for *L. intracellularis*. Expected size for 16S RNA and *aspA* is 250 bp and 141 bp, respectively. Orange arrows indicate non-specific PCR amplification products.

Table 2.2 Sequencing statistics of samples processed after 16S rRNA based quantification.

Sample name	Sample type	16s rRNA copy per ng of genomic DNA	Raw reads generated	Percentage of reads mapping to host	Percentage of reads mapping to *L. intracellularis* [a]	Mean *L. intracellularis* depth of coverage
DKp23	Cell cultured	1.91×10^6	82,879,784	85	97	812×
5189	Cell cultured	1.86×10^6	80,104,306	10	3.7	182×
916/91[b]	Cell cultured	4.79×10^5	75,720,948	48	0.1	0.7×
15540	Cell cultured	3.50×10^5	80,685,770	89	98	603×
1482/89p19	Cell cultured	1.49×10^5	76,623,876	59	0.2	3×
1482/89p20	Cell cultured	1.24×10^5	76,316,872	68	0.5	10×
Edinburgh	Colon tissue	2.68×10^5	86,147,194	99	43	34×
4241	Ileum tissue	1.86×10^5	83,521,302	1.4	12	7×
4242	Ileum tissue	1.43×10^4	78,767,308	95	19	14×
6073	Faecal	7.48×10^5	61,746,046	0.8	0.2	30×
3387	Faecal	5.37×10^5	86,147,194	0.23	0.2	8×
2069	Faecal	1.90×10^5	80,014,282	0.4	1	67×
1383	Faecal	9.20×10^4	98,797,652	99	0.1	0.7×

[a] Percentage of reads mapping to reference genome PHE/MN1-00 after host reads removal

[b] Mycoplasma contamination in sample

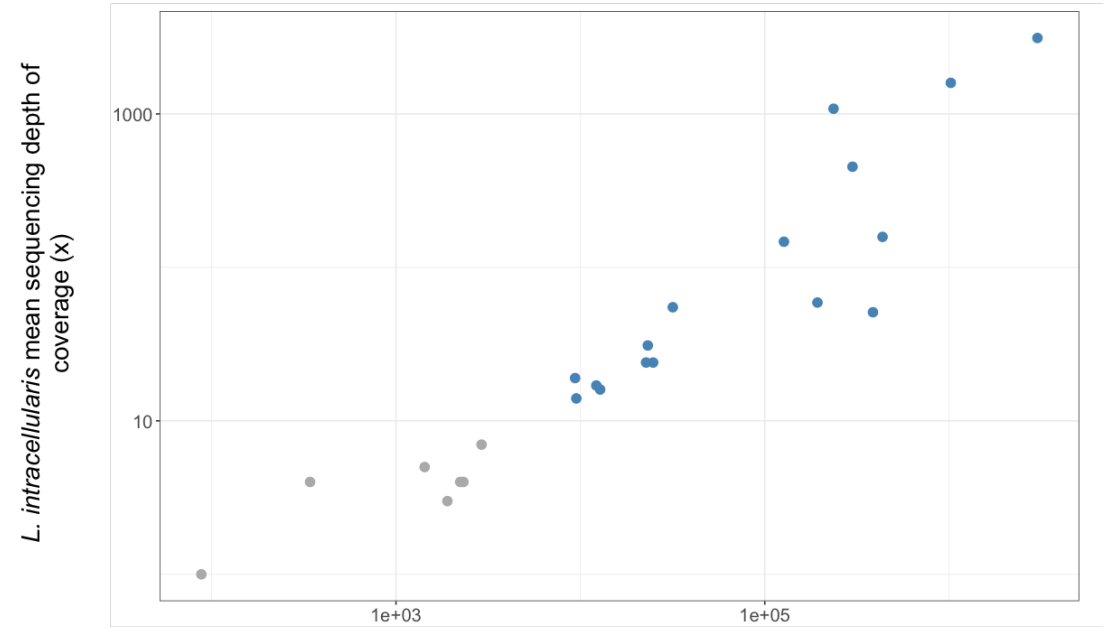
2.4.2 Sequencing analysis of clinical samples

The minimum sequencing depth required for genome assembly is at approximately 20× (Desai et al., 2013). Thus, to establish the minimum copy number of *L. intracellularis* DNA required within a sample in order to obtain sequencing depth of ≥20×, we compared the sequencing depth and genome coverage generated in relation to the quantity of *L. intracellularis* DNA in the sequencing library. A total of 16 porcine and 7 equine faecal samples with varying quantities of *L. intracellularis* DNA, measured using qPCR targeting *aspA* were submitted for sequencing. Multiple libraries were pooled and multiplexed on a single flow cell lane generating between 55 and 243 million reads per sample (Table 2.3). We observed a direct correlation between *aspA* copy number and sequencing depth of *L. intracellularis* (Figure 2.2A). Furthermore, samples with *aspA* copy numbers > 9 $\times 10^4$ per ng of genomic DNA achieved genome coverage of > 98% across the reference genome (Figure 2.2B). The two percent genome coverage absent mapped to an 18 kb prophage region strain dependent for PHE/MN1-00.

Faecal sample 8163, containing the most abundant amount of *L. intracellularis* DNA at 3.01×10^6 *aspA* copy number per ng of genomic DNA, generated a sequencing depth of 3,125× and genome coverage of 98%, with 39% reads assigned to *L. intracellularis* (Table 2.3). Faecal samples containing copy number of *aspA* at 2.27×10^4 per ng of genomic DNA achieved ≥23× sequencing depth and >98% coverage across the reference genome PHE/MN1-00, with 0.12% reads assigned to *L. intracellularis* (Table 2.3). Majority of faecal samples contained large percentage of reads (ranging from 50 to 95%) that remained unclassified when mapped against the RefSeq database containing bacterial, archaeal, viral domains and human genome from NCBI (Table 2.3).

Overall, we established a minimum of 2 $\times 10^4$ *L. intracellularis* DNA copy number per ng of genomic DNA, is required to achieve sufficient sequencing depth and coverage for genome assembly and variant calling. Of the 23 faecal samples sequenced, 12 porcine samples met the requirement.

(A)


L. intracellularis mean sequencing depth of coverage (x)
1000
10
1e+03
1e+05
aspA copy numbers per ng of genomic DNA

(B)

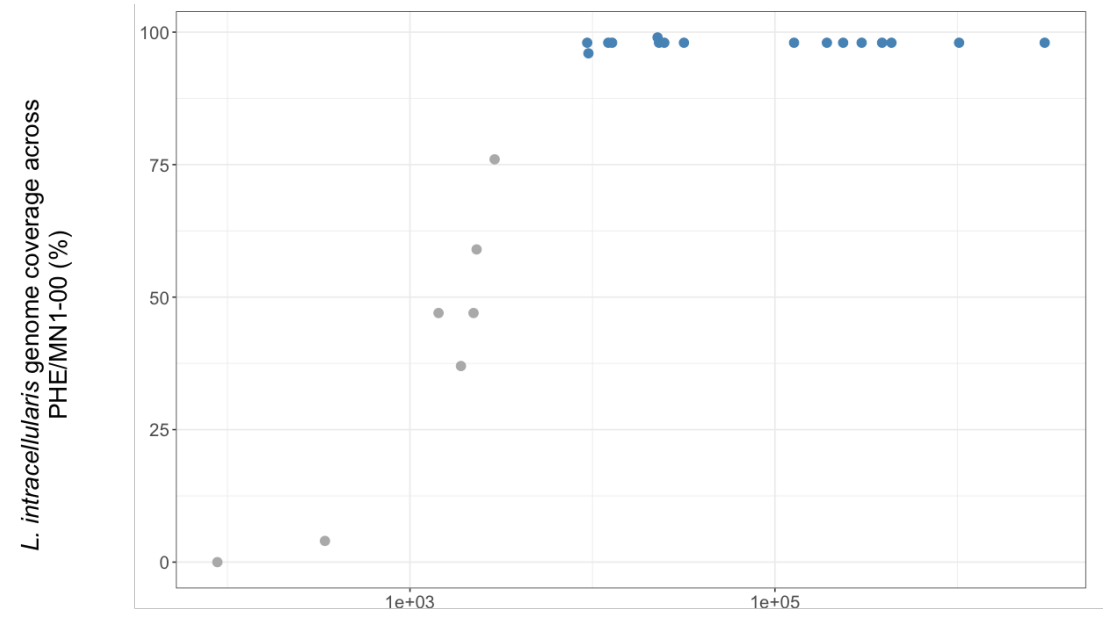

L. intracellularis genome coverage across PHE/MN1-00 (%)
100
75
50
25
0
1e+03
1e+05
aspA copy numbers per ng of genomic DNA

Figure 2.2 Correlation between *L. intracellularis aspA* copy numbers in sequencing library and sequencing outcome. (A) *L. intracell*ularis DNA copy number from faecal samples were quantified by measuring *aspA* copy number per ng of genomic DNA in sequencing library, this was proportional to the mean bacterium sequencing depth of coverage. Axis are in log10 scale. Grey dots represent samples which resulted in a sequencing depth of less than 20×, blue dots represent samples with a sequencing depth of more than 20× that were sufficient for de novo assembly. (B) *aspA* copy numbers and correlation to percentage of coverage across *L. intracellularis* reference genome PHE/MN1-00. X-axis is in log10 scale. Grey dots represent samples achieved less than 98% coverage across the reference genome.

Table 2.3 Sequencing statistics of faecal DNA samples.

Sample ID	Host	*aspA* copy per ng of genomic DNA	Raw reads generated	Percentage of reads mapping to host	Percentage of reads mapping to *L. intracellularis* [a]	Mean *L. intracellularis* depth of coverage	Percentage of unclassified reads [b]
8163	Porcine	3.01×10^6	122,364,098	3.31	39.15	3,125×	50
SRUC1	Porcine	1.02×10^6	126,914,078	29.75	22.07	1,594×	19
682	Porcine	4.35×10^5	116,522,970	11.84	1.40	158×	94
630	Porcine	3.86×10^5	140,409,176	9.19	0.8	51×	92
5939	Porcine	2.99×10^5	146,304,454	0.14	4.81	453×	85
661	Porcine	2.36×10^5	154,170,554	0.2	11.1	1,081×	85
1886	Porcine	1.93×10^5	55,160,168	0.17	1.74	59×	87
976/1	Porcine	1.27×10^5	153,251,638	0.12	1.49	147×	85
2746	Porcine	3.17×10^4	183,677,484	0.09	0.42	55×	85
5626	Porcine	12.48×10^4	135,379,604	0.10	0.38	24×	92
6073	Porcine	2.32×10^4	207,677,394	0.43	0.33	30×	90
F22	Porcine	2.27×10^4	97,141,250	2.18	0.42	23×	95

[a] Percentage of reads mapping to reference genome PHE/MN1-00 after host reads removal

[b] Percentage of unclassified reads mapping to standard kraken 2 database

Table 2.3. Sequencing statistics of faecal DNA samples (continued)

Sample ID	Host	*aspA* copy per ng of genomic DNA	Raw reads generated	Percentage of reads mapping to host	Percentage of reads mapping to *L. intracellularis* [a]	Mean depth of *L. intracellularis* coverage	Percentage of unclassified reads [b]
3387	Porcine	9.50×10^3	117,175,932	0.12	0.15	14×	82
SRUC3	Porcine	9.36×10^3	227,380,016	0.07	0.12	19×	90
1660	Porcine	2.91×10^3	170,137,008	0.09	0.17	7×	80
5211	Porcine	2.23×10^3	119,232,752	0.14	0.09	4×	88
H9	Equine	1.28×10^4	115,280,262	0.39	0.29	16×	91
H14	Equine	1.22×10^4	163,200,202	0.39	0.26	17×	52
H24	Equine	2.32×10^3	130,189,090	0.10	0.11	4×	86
H8	Equine	1.90×10^3	155,755,644	0.08	0.11	3×	88
H5	Equine	1.43×10^3	243,472,874	0.05	0.17	5×	72
H21	Equine	3.42×10^2	195,172,690	0.10	0.10	3×	91
H16	Equine	8.80×10^1	213,167,694	0.06	0.09	1×	72

[a] Percentage of reads mapping to reference genome PHE/MN1-00 after host reads removal

[b] Percentage of unclassified reads mapping to standard kraken 2 database

2.4.3 Microbial DNA enrichment maximised sequencing depth of coverage from tissue samples

In order to maximise sequencing of *L. intracellularis* through direct sequencing of clinical samples, we assessed the effectiveness of the NEBNext Microbiome DNA Enrichment kit to enhance *L. intracellularis* sequencing depth. This enrichment kit utilises magnetic beads coated with recombinant proteins consisting of methyl-CpG-binding domain protein 2 (MBD2) fused to a Fc tail of human IgG (Feehery et al., 2013). The MBD2 is specific for DNA which is methylated at position 5 of cytosine. In human and other eukaryotic DNA, this modification is often found in CpG dinucleotides and occurs at a higher rate than in prokaryotes. On the basis of differences in CpG methylation abundance, the beads are able to selectively separate prokaryotic DNA from vertebrate DNA within samples (Feehery et al., 2013). Consequently, the depletion of host DNA from samples will enrich for microbial DNA yield for sequencing.

Here, the kit was used for microbial DNA enrichment from a tissue and a faecal sample. We compared and evaluated the sequencing outcome of the DNA samples before and after the enrichment step. Without application of the enrichment kit, ileal sample 630 with 2.44×10^2 *aspA* per ng of genomic DNA achieved *L. intracellularis* sequencing depth of 49× and 38% reads assigned to the bacterium (Table 2.4). Following application of the enrichment step, ileal sample with 1.65×10^2 *aspA* per ng of genomic DNA achieved sequencing depth of 180× and 67% reads assigned to *L. intracellularis* (Table 2.4). Thus, inclusion of the enrichment step improved sequencing depth of coverage by 3-fold (Figure 2.3).

For the faecal sample 682, we observed a decrease of sequencing depth by 36× following enrichment (Figure 2.4). While sample without application of the enrichment kit, containing 3.47×10^3 *aspA* per ng of genomic DNA achieved *L. intracellularis* sequencing depth of 159× and 1.46% reads assigned to the bacterium (Table 2.4). After enrichment, sample with 1.00×10^3 *aspA* per ng of genomic DNA achieved a sequencing depth of 120× and 1.92% reads assigned to *L. intracellularis* (Table 2.4). We noticed that there was a decrease in the number of raw reads generated for the enriched faecal DNA sample compared to the non-enriched, from approximately 220 to 167 million, respectively (Table 2.4). Therefore, the decrease in *L. intracellularis* sequencing depth in the enriched sample could partially be due to a lower number of raw reads generated.

For both sample types, a decrease in *L. intracellularis* copy number was observed, after treatment with the enrichment kit (Table 2.4). PCR detection of DNA that remained bound to the bead complex and within the wash buffer revealed the presence of *L. intracellularis* DNA (Figure 2.5). This suggests non-specific binding of prokaryotic DNA to the MBD2 recombinant proteins. Although, higher levels of non-specific binding were observed in the enrichment of faecal sample than compared to the ileal sample (Figure 2.5B).

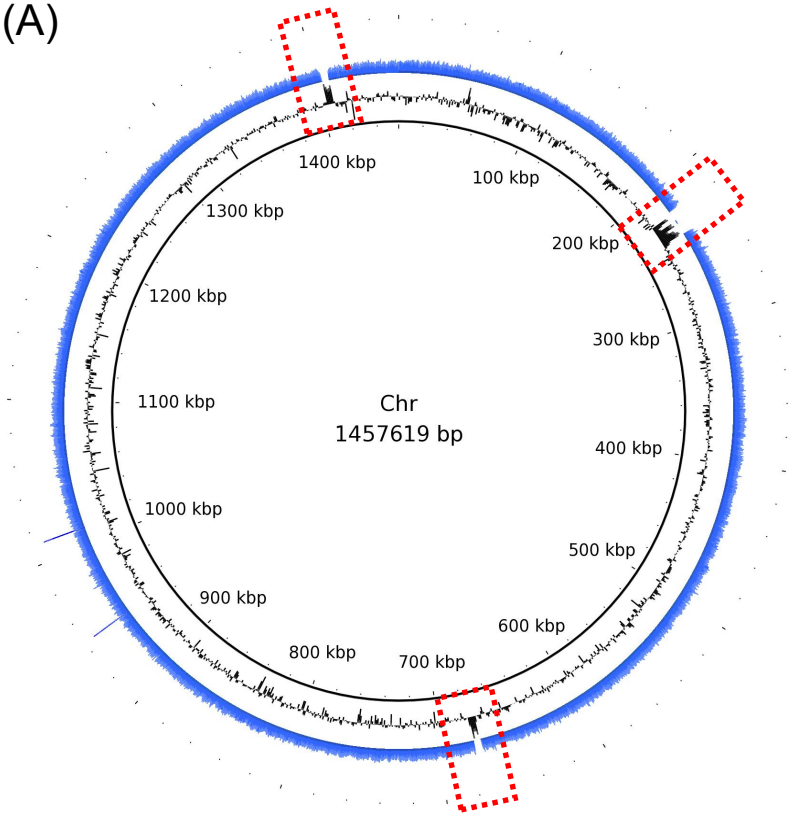
Overall, the NEBNext Microbiome DNA Enrichment kit was effective for the enrichment of microbial DNA in the ileal sample, maximising the sequencing depth of *L. intracellularis*. We then used the enrichment kit to facilitate the sequencing of two more ileal samples, Thirsk2 and 4242 (Table 2.4).

Table 2.4 Sequencing statistics of enriched and non-enriched DNA samples.

Sample ID	Microbial DNA enrichment	Sample type	*aspA* copy per ng of genomic DNA	Raw reads generated	Percentage of reads mapping to host	Percentage of reads mapping to *L. intracellularis* [a]	Mean *L. intracellularis* depth of coverage
630	No	Ileal	2.44×10^2	156,779,116	98	38	49×
630	Yes	Ileal	1.65×10^2	171,062,260	97	67	180×
682	No	Faecal	3.47×10^3	222,924,454	12	1.40	156×
682	Yes	Faecal	1.00×10^3	167,149,180	5	1.92	120×
Thirsk2	Yes	Ileal	6.13×10^3	92,510,154	94	6.35	21×
4242	Yes	Ileal	3.04×10^3	92,785,530	99	38.06	18×

[a] Percentage of reads mapping to reference genome PHE/MN1-00 after host reads removal

(A)


Chr
1457619 bp
100 kbp
200 kbp
300 kbp
400 kbp
500 kbp
600 kbp
700 kbp
800 kbp
900 kbp
1000 kbp
1100 kbp
1200 kbp
1300 kbp
1400 kbp

(B)

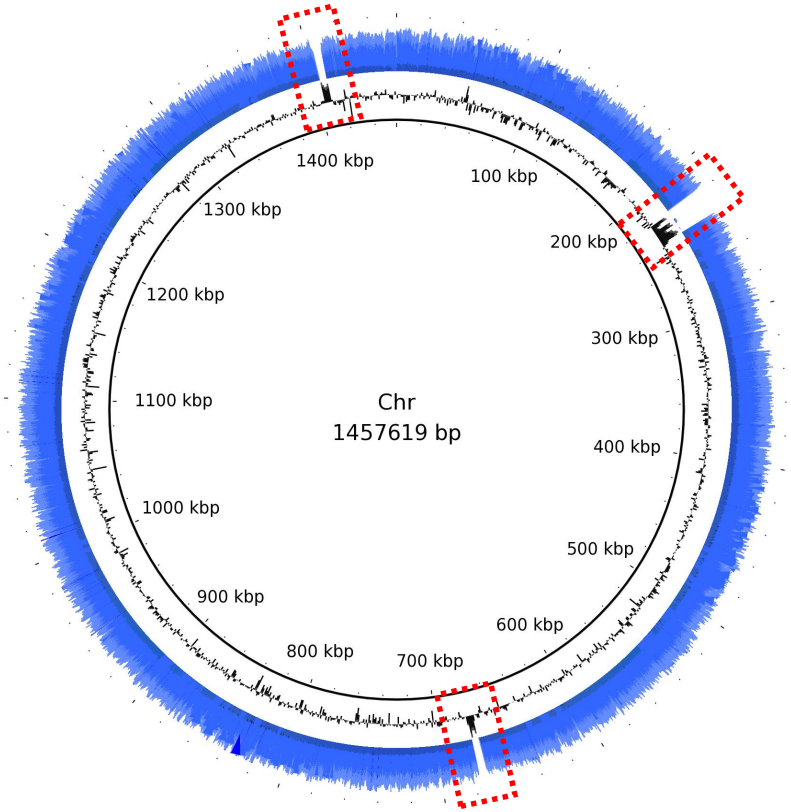
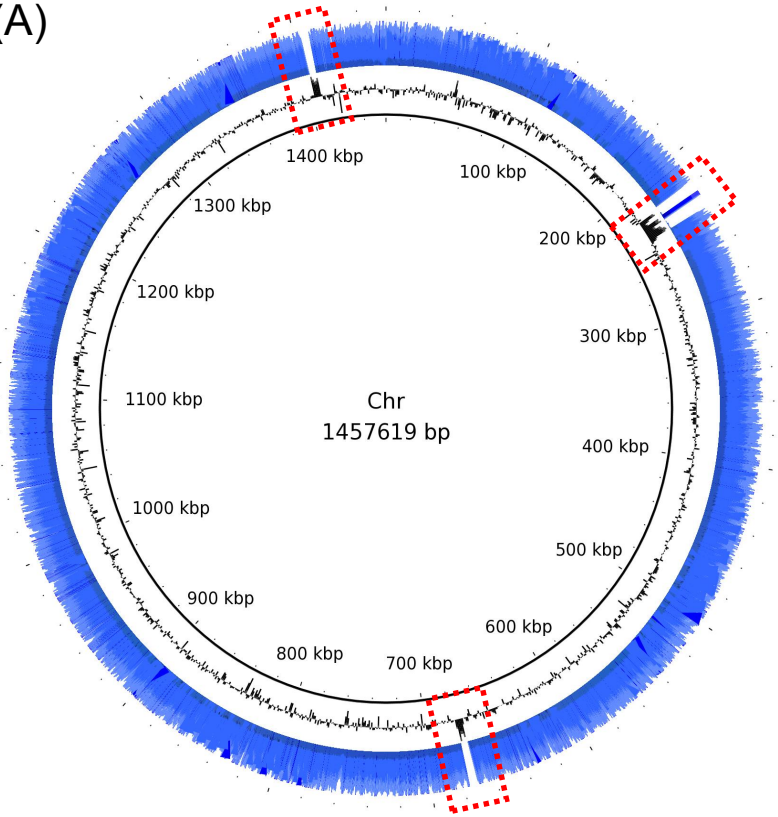

Chr
1457619 bp
100 kbp
200 kbp
300 kbp
400 kbp
500 kbp
600 kbp
700 kbp
800 kbp
900 kbp
1000 kbp
1100 kbp
1200 kbp
1300 kbp
1400 kbp

Figure 2.3 Assessment of microbial enrichment to enhance *L. intracellularis* sequencing from ileal DNA. Read mapping coverage from sequencing of (A) ileal DNA without and (B) with enrichment treatment using the NEBNext Microbiome DNA Enrichement Kit. Sequencing reads are mapped against the chromosome of *L. intracellularis* PHE/MN1-00 reference genome. Middle circle represents GC content across the reference genome, blue outer ring represents mapping coverage of *L. intracellularis* reads to the reference, by which the thickness of the blue ring represents depth of coverage. Following enrichment treatment of ileal DNA sequencing depth of *L. intracellularis* increased by three-fold. Three regions without sequence coverage across the four rings are highlighted in red boxes, these map to two copies of the 16S rRNA gene and the prophage-associated genomic island.

(A)


Chr
1457619 bp
100 kbp
200 kbp
300 kbp
400 kbp
500 kbp
600 kbp
700 kbp
800 kbp
900 kbp
1000 kbp
1100 kbp
1200 kbp
1300 kbp
1400 kbp

(B)

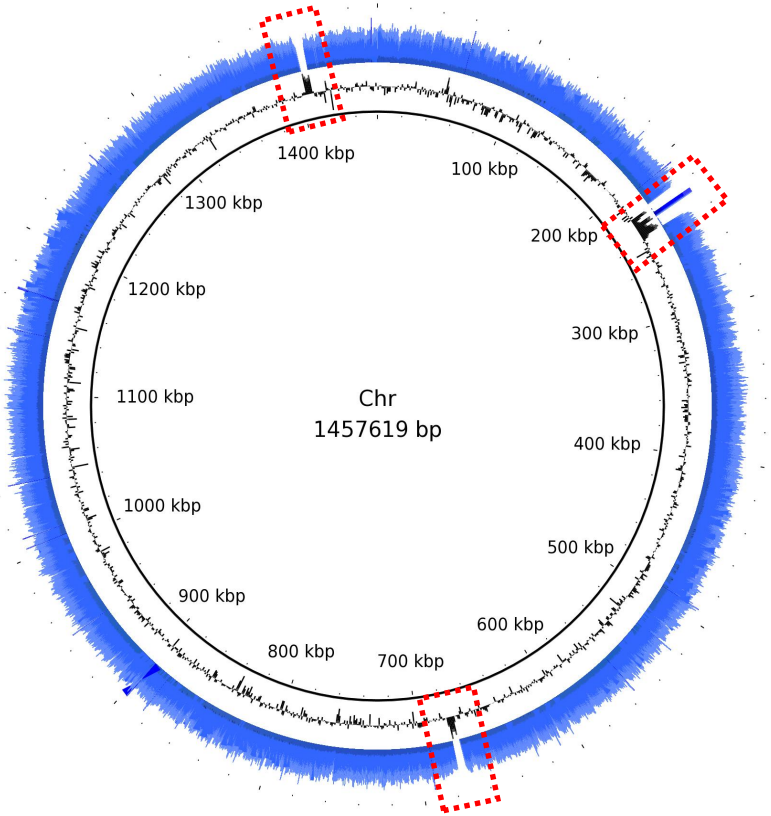

Chr
1457619 bp
100 kbp
200 kbp
300 kbp
400 kbp
500 kbp
600 kbp
700 kbp
800 kbp
900 kbp
1000 kbp
1100 kbp
1200 kbp
1300 kbp
1400 kbp

Figure 2.4 Assessment of microbial enrichment to enhance *L. intracellularis* sequencing from faecal DNA. coverage of sequencing reads from (A) faecal DNA without and (B) with enrichment treatment using the NEBNext Microbiome DNA Enrichement Kit. Sequencing reads are mapped against the chromosome of *L. intracellularis* PHE/MN1-00 reference genome. Sequencing depth of *L. intracellularis* from faecal DNA decreased from 156× to 120× following enrichment treatment.

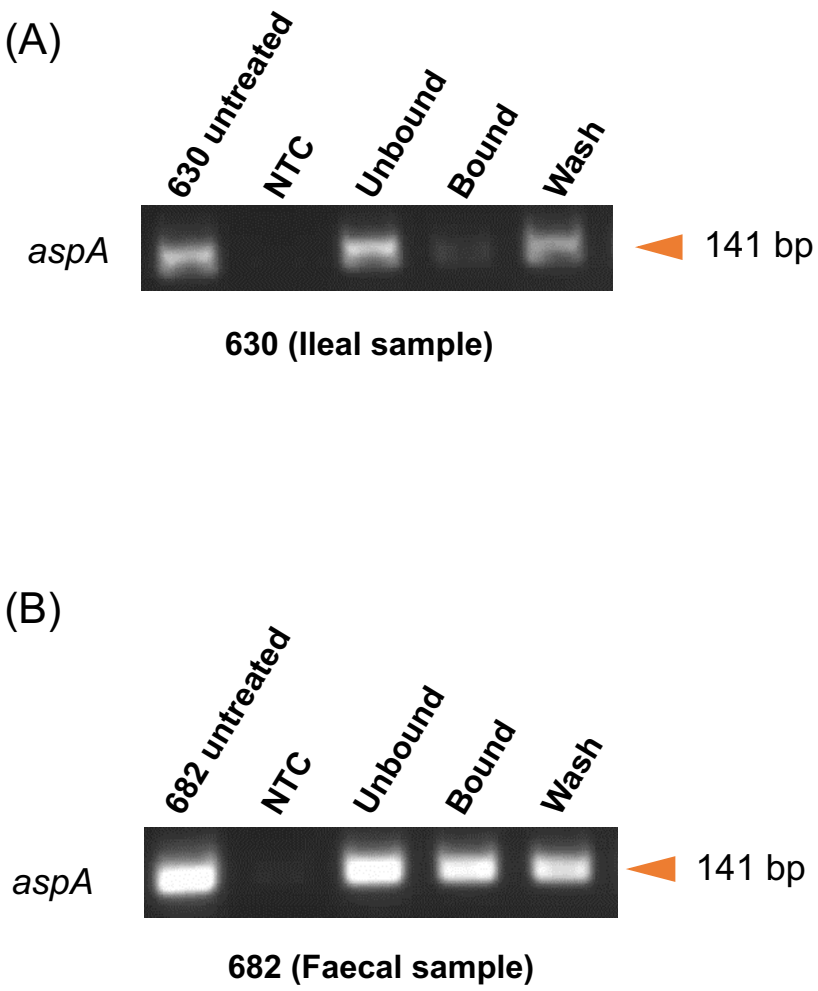


Figure 2.5 Detection of *L. intracellularis* DNA from different enrichment fractions. PCR targeting the *aspA* gene to detect presence of *L. intracellularis* DNA in different enrichment fractions from treatment of (A) ileal DNA and (B) faecal DNA. The untreated samples represent DNA without treatment with the NEBNext Microbiome DNA Enrichment kit. "Unbound" corresponds to the unbound DNA remaining in the supernatant after host DNA removal with magnetic beads, "Bound" corresponds to the DNA remained bound to magnetic beads, and "Wash" corresponds to DNA in the supernatant used for washing of the beads.

2.4.4 Assessment of metagenomic assemblers on *L. intracellularis* draft genome reconstruction

Three of the most widely used metagenomic assemblers, metaSPAdes, MEGAHIT and IDBA-UD, are DBG based and have been developed for the assembly of sequencing data with highly nonuniform depth of coverage (Vollmers et al., 2017). We benchmarked the three assemblers on five faecal DNA samples to investigate which assembler is the most appropriate for our dataset. Here, we solely focused on the assembly quality of *L. intracellularis*. Hence, the five samples were selected based on having different *L. intracellularis* depth of coverage, ranging from 158× to 7× (Table 2.5), to determine the limit of genome assembly.

The quality of *L. intracellularis* genome assembly from the three assemblers was assessed using QUAST, which aligns the assembled contigs to a reference genome and computes several metrics relative to the reference (Gurevich et al., 2013). These include; i) the percentage of reference genome coverage, ii) the minimum contig length that comprise over half of the entire assembly (N_{50} size), iii) the maximum contig length, iv) the number of misassemblies and v) misassembled contig lengths. Here, *L. intracellularis* PHE/MN1-00 reference genome was used to assess genome assembly quality. Our assembly results demonstrated that generally, a decrease in *L. intracellularis* depth of coverage resulted in genome assembly becoming more fragmented, as indicated by a decrease in N_{50} values (Table 2.5). Genome assembly with metaSPAdes produced better contiguity, with larger N_{50} value and longer contig length (Table 2.5). However, it appears that metaSPAde also produced the longest misassembled contig length (Table 2.5). Although bearing in mind this metric is based on the length of contigs containing the misassemblies (Gurevich et al., 2013). Thus,

the more contiguous assembly generated by metaSPAde may lead to an increase in the misassembled contig length. Assembly with MEGAHIT generally produced a slightly larger total genome size and achieved the highest recovery on low-coverage subset, by which assembly of 1660 with *L. intracellularis* sequence depth of 7× produced genome fraction of 86.5% (Table 2.5).

In addition to quality assessment, we evaluated the assembly cost and efficiency of the three assemblers, measuring the total memory and runtime used. Performance wise, metaSPAde required the highest computational cost and runtime, by which assembly of sample 682 required 100 GB and ran for 24 hrs (Table 2.6). MEGAHIT was the most efficient for memory usage, in which assembly of sample 682 only required 18 GB, almost six times less than metaSPAde (Table 2.6). In terms of speed, IDBA-UD performed the best and took approximately half the time of MEGAHIT to complete assembly. For sample 3387, IDBA-UD performed six times faster than metaSPAde (Table 2.6). We noticed that there was no correlation between the size of the file and the memory required for its assembly, by which a file of 52 GB (sample 682) required half the amount of memory to assemble than compared to a file of 43 GB (sample 630).

Overall, metaSPAde assembler performed the best on the basis of producing the most contiguous assembly with minimal number of misassemblies. However, assembly using MEGAHIT had a good compromise between quality of genome assembly and assembly efficiency.

Table 2.5 Comparison of assembly quality using different metagenomic assembler.

Sample	*L. intracellularis* depth of coverage	Assembler	Genome fraction (%) [a]	N_{50} (bp)	Largest contig (bp)	Total (bp)	Misassemblies	Misassembled contigs length (bp)
682	158×	metaSPAdes	98.6	391,926	552,536	1,695,329	5	757,502
		MEGAHIT	98.7	180,477	421,141	1,694,999	6	486,805
		IDBA-UD	98.6	391,970	601,569	1694,933	5	857,295
630	51×	metaSPAdes	98.6	335,197	552,514	1,695,213	4	753,349
		MEGAHIT	98.7	228,756	421,144	1,697,152	3	344,212
		IDBA-UD	98.6	159,727	408,506	1,695,241	5	333,476
H9	16×	metaSPAdes	97.6	340,232	585,655	1,689,674	3	165,385
		MEGAHIT	97.7	132,840	303,080	1,743,070	4	198,171
		IDBA-UD	97.8	179,813	473,353	1,655,846	3	91,853
3387	14×	metaSPAdes	98.4	60,837	183,447	1,694,564	5	265,821
		MEGAHIT	98.5	60,131	133,054	1,695,663	5	227,146
		IDBA-UD	98.3	43,283	149,047	1,689,743	5	161,916

[a] Percentage of assembled genome coverage across the reference genome PHE/MN1-00

Table 2.5 Comparison of assembly quality using different metagenomic assembler (continue)

Sample	*L. intracellularis* depth of coverage	Assembler	Genome fraction (%) [a]	N_{50} (bp)	Largest contig (bp)	Total (bp)	Misassemblies	Misassembled contigs length (bp)
1660	7×	metaSPAdes	72.3	2,713	18,659	1,245,203	3	26,496
		MEGAHIT	86.5	1,921	8,113	1,492,701	2	4,552
		IDBA-UD	60	1,206	8,032	1,083,130	0	0

[a] Percentage of assembled genome coverage across the reference genome PHE/MN1-00

Table 2.6 Performance comparison of metagenomic assemblers.

Sample	Sample size	Input reads	Assembler	Memory required	Run time (hours)
682	52 GB	79,823,201	metaSPAdes	100GB	24
			MEGAHIT	18 GB	17
			IDBA-UD	58 GB	8
1660	43 GB	64,614,260	metaSPAdes	236 GB	62
			MEGAHIT	24 GB	28
			IDBA-UD	138 GB	23
3387	36 GB	48,694,236	metaSPAdes	170 GB	68
			MEGAHIT	18 GB	25
			IDBA-UD	74 GB	11
H9	31 GB	45,408,569	metaSPAdes	106 GB	19
			MEGAHIT	16 GB	12.5
			IDBA-UD	58 GB	7

Table 2.6 Performance comparison of metagenomic assemblers (continue)

Sample	Sample size	Input reads	Assembler	Memory required	Run time (hours)
630	24 GB	44,108,824	metaSPAdes	99 GB	34
			MEGAHIT	19 GB	17
			IDBA-UD	56 GB	6

2.4.5 Unsupervised binning unable to recover *L. intracellularis* plasmid 1 and plasmid 2

In order to reduce the reliance of using a reference genome, we evaluated the efficiency of unsupervised binning using metaBAT, to recover *L. intracellularis* draft genome from MAGs (Kang et al., 2015). The quality of genome bins produced by metaBAT was assessed based on the presence and absence of lineage-specific single-copy marker genes by checkm, as an estimation of completeness and contamination of the genome (Parks et al., 2015). metaBAT software measures the probabilistic distance of tetra-nucleotide frequency (TNF) between a contig pair, then determines whether the contig pair originates from the same genome or not using a precomputed posterior probability distribution model, calculated based on the inter and intra Euclidean TNF distance of 1,414 reference genomes (Kang et al., 2015). Difference in mean base coverage of the contig pair is also measured. The TNF and mean coverage distance probability of each contig pair are combined and a distance matrix between all pairs of contigs are calculated and used for clustering, with the assumption that contigs originating from the same genome should share similar TNF and sequence coverage (Kang et al., 2015).

Here, we only focused on the quality of *L. intracellularis* genome bin from each metagenome assembly. Results revealed that metaBAT was able to produce genome bins with >90% completeness for four samples (Figure 2.6 and Table 2.7). Manual inspection of 682, 630, H9 and 3387 *L. intracellularis* genome bins revealed that the ~10% genome missing represented absence of plasmid 1, plasmid 2 and the two copies of 16S, 23S and 5S rRNA genes (Figure 2.7). Ten percent contamination was observed in the genome bin of sample H9, by which inspection of the contaminant

contigs demonstrated the presence of genomic DNA fragments ranging from 3 to 45 kb in length, from various bacterial species (data not shown). Inspection of contig alignment within each genome bin against the PHE/MN1-00 reference genome revealed the absence of an approximately 50 kb region in sample 682 and 3387 (Figure 2.6), which did not impact checkm estimation of completeness (Figure 2.5 and Table 2.7). Sample 1660 with the lowest N_{50} assembly value produced the lowest quality genome bin with 50% completeness (Figure 2.5 and Table 2.7). Generally, poor quality metagenome assembly will result in generation of poor-quality genome bins, due to the fact that a reduction in contig size will lead to a decrease in metaBAT ability to discriminate TNF distance between the contigs.

Overall, unsupervised binning through metaBAT was able to produce *L. intracellularis* genome bins of >90% completeness, given a good quality metagenome assembly. However, recovery of plasmids appeared to be challenging and a reference genome is still required to assist the recovery of *L. intracellularis* genome.

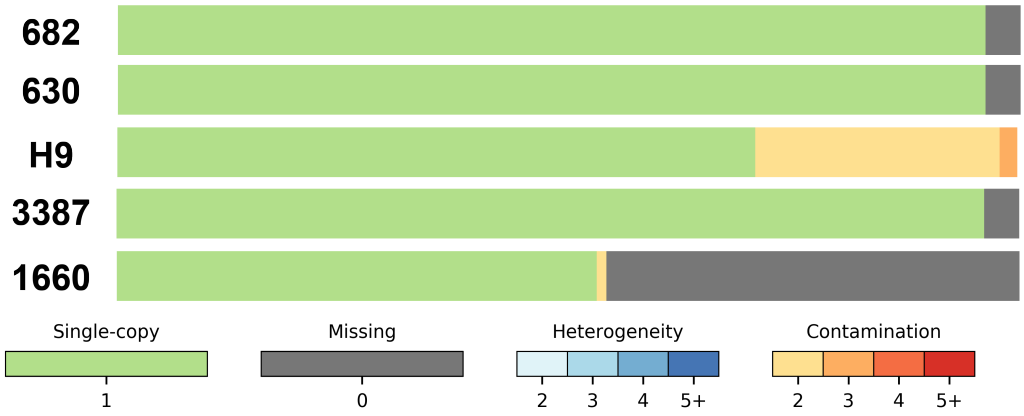


Figure 2.6 Quality assessment plot of *L. intracellularis* genome bins. The green bar represents completeness of the bin measured by the presence of single markers identified once in the bin. Percentage of completeness and contamination are described in Table 2.8. The grey bar represents single marker genes that are missing from the bin. Different shades of blue or red bars represent single markers identified multiple times (2-5+) in each genome bin, indicative of contamination. Pairs of multi-copy genes with amino acid identity (AAI) ≥90% are indicated in blue and genes with less than 90% are indicated in red.

Table 2.7 Table describing completeness and contamination of each *L. intracellularis* genome bin.

Sample ID	N_{50}	Unique markers (of 43)	Multi-copy	Completeness	Contamination
682	391,926	43	0	93.49	0.00
630	335,197	43	0	93.49	0.00
H9	340,232	22	21	100.00	10.53
3387	60,837	43	0	93.49	0.00
1660	2,713	31	1	50.35	0.07

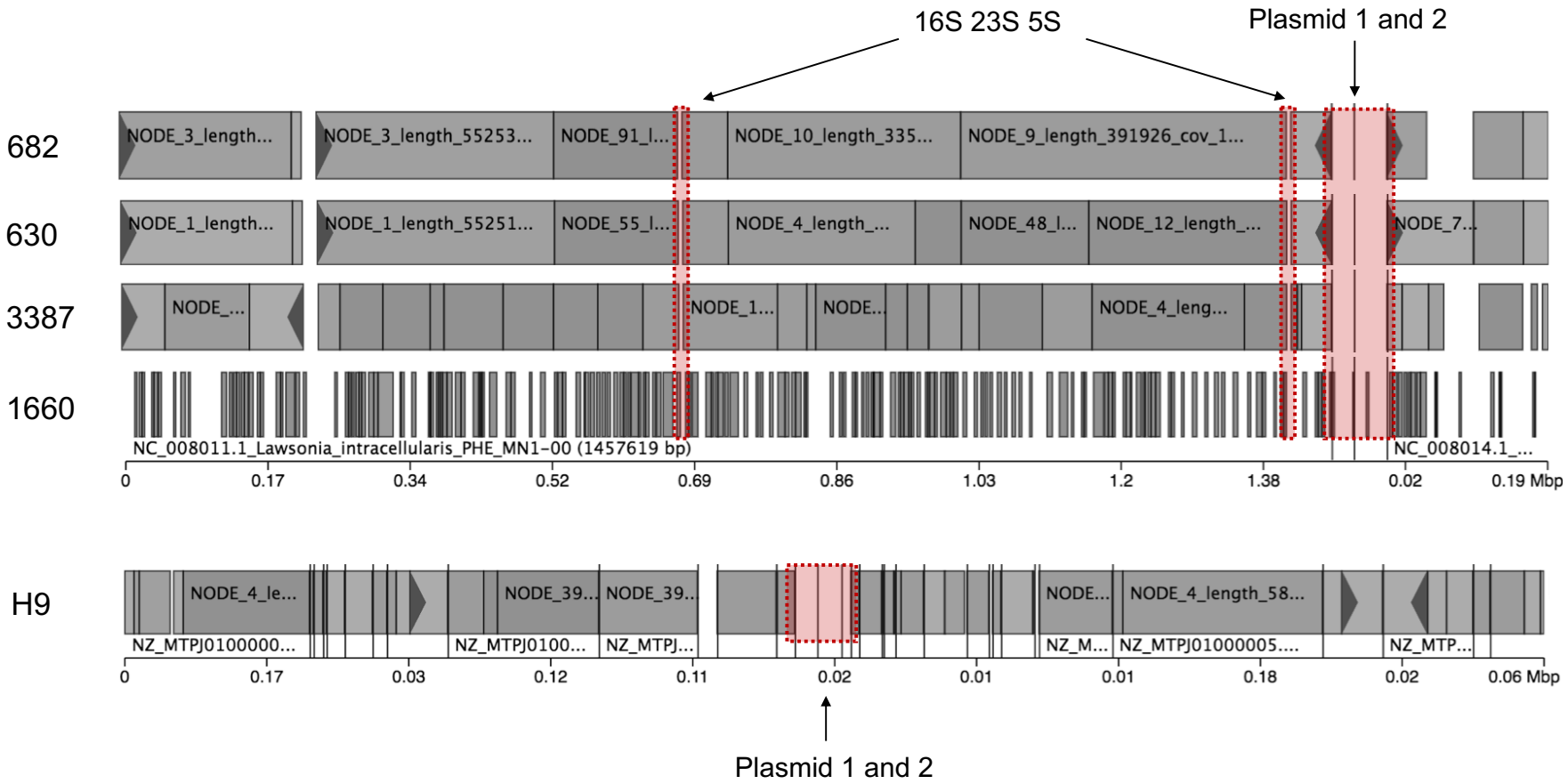

16S 23S 5S
Plasmid 1 and 2
682
630
3387
1660
H9
NODE_3_length...
NODE_3_length_55253...
NODE_91_l...
NODE_10_length_335...
NODE_9_length_391926_cov_1...
NODE_1_length...
NODE_1_length_55251...
NODE_55_l...
NODE_4_length_...
NODE_48_l...
NODE_12_length_...
NODE_7...
NODE_...
NODE_1...
NODE...
NODE_4_leng...
NC_008011.1_Lawsonia_intracellularis_PHE_MN1-00 (1457619 bp)
NC_008014.1_...
0
0.17
0.34
0.52
0.69
0.86
1.03
1.2
1.38
0.02
0.19 Mbp
NODE_4_le...
NODE_39...
NODE_39..
NODE...
NODE_4_length_58...
NZ_MTPJ0100000...
NZ_MTPJ0100...
NZ_MTPJ...
NZ_M...
NZ_MTPJ01000005....
NZ_MTP...
0
0.17
0.03
0.12
0.11
0.02
0.01
0.01
0.18
0.02
0.06 Mbp
Plasmid 1 and 2

Figure 2.7 *L. intracellularis* genome bins. Contigs placed in *L. intracellularis* genome bin from sample 682, 630, 3387 and 1660 were mapped against porcine associated strain PHE/MN1-00 reference genome which all revealed absence of the two copies of 16S, 23S and 5S rRNA genes, plasmid 1 and plasmid 2. Contigs placed in *L. intracellularis* genome bin of H9 were mapped against equine associated strain E40504 reference genome, by which plasmid 1 and plasmid 2 were also missing.

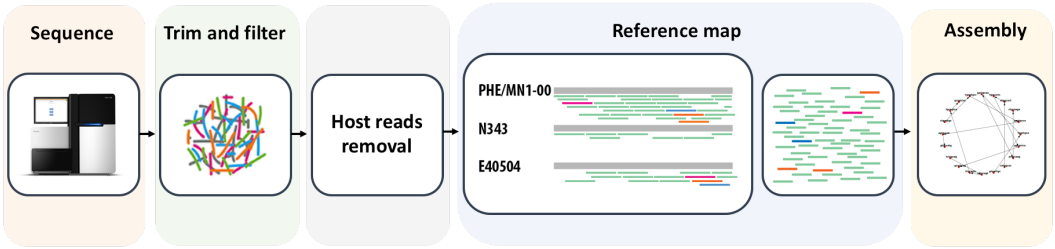
2.4.6 A reference-guided *de novo* assembly approach improves draft genome reconstruction

In a further attempt to improve genome assembly quality, a reference-guided *de novo* assembly strategy was adapted (Figure 2.7). This involves extracting *L. intracellularis* reads from the metagenomic dataset prior to *de novo* assembly, in order to reduce the complexity of the data. Kraken was used to extract *L. intracellularis* reads, by which a custom classification tree was constructed using a database containing three *L. intracellularis* reference genomes, E40504 (GenBank GCA_001975945.1), N343 (GenBank GCA_000331715.1) and PHE/MN1-00 (GenBank GCA_000055945.1). Multiple references were included in the classification tree to capture as much diversity as possible and reduce bias towards a single reference. Classified reads that mapped to the tree were then used as input for genome assembly with metaSPAde.

Comparison of *L. intracellularis* genome assembly quality with reference-guided and reference independent approach demonstrated that the former generally improved assembly contiguity, generating larger N_{50} values and larger contigs length (Table 2.8). We also observed a decrease in the number of misassemblies generated using the reference-guided approach (Table 2.8). Furthermore, an improvement in genome fraction from 72 to 92% was observed in assembly of sample 1660, suggesting the reference-guided approach is able to improve assembly of low coverage dataset (Table 2.8). Comparison of the assemblies produced by the two different strategies revealed minor variation in the number of CDS and total assembly length, which ensured consistency between the two approaches (Table 2.8). Furthermore, using this approach we were able to produce genome assembly with more than 98% genome fraction for sample 3387, containing *L. intracellularis* sequencing depth 14× and generating only two misassemblies (Table 2.8).

Overall, using the reference-guided assembly we were able to reconstruct 21 *L. intracellularis* draft genomes with more than 97% genome fraction (Table 2.9), from a total of 30 metagenomic datasets sequenced from tissue and faecal samples.

(A)



(B)

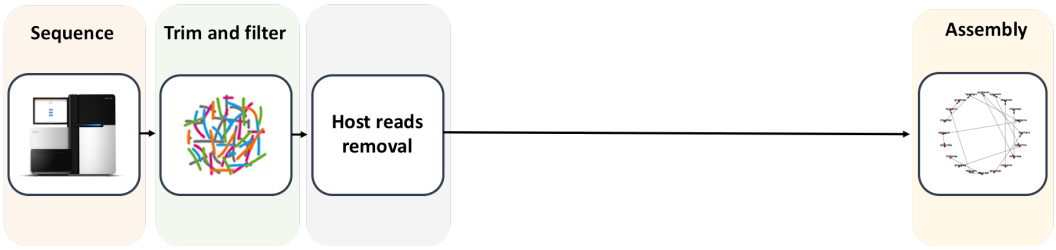


Figure 2.8 Flow chart outlining the pipeline for reference-guided and reference independent de novo assembly. (A) Reference-guided assembly requires mapping of reads to a reference database prior to de novo assembly in order to reduce the complexity of the data. Following host reads removal, the remaining reads were mapped against a kraken database containing *L. intracellularis* reference genomes, during which sequences such as 16S rRNA gene may also be extracted as it is highly conserved across bacteria. (B) Reference independent assembly do not rely on availability of reference genome and is performed directly on host filtered reads, followed by binning of assembled contigs to recover *L. intracellularis* genome from MAGs.

Table 2.8 Comparison of assembly statistics between reference-guided and reference independent assembly.

Sample	Depth of coverage [a]	Assembly method	Genome fraction (%) [b]	N_{50} (bp)	Largest contig (bp)	Total (bp)	Misassemblies	Misassembled contigs length (bp)	No. CDS
682	158×	Reference-guided	98.7	391,994	701,458	1,696,555	1	701,458	1,420
		Reference independent	98.6	391,926	552,536	1,695,329	5	757,502	1,422
630	51×	Reference-guided	98.7	395,996	701,725	1,698,369	1	701,725	1,420
		Reference independent	98.6	335,197	552,514	1,695,213	4	753,349	1,423
H9	16×	Reference-guided	97.8	217,170	328,180	1,657,068	2	395,954	1,416
		Reference independent	97.6	340,232	585,655	1,689,674	3	165,385	1,412

[a] *L. intracellularis* depth of coverage, [b] Percentage of coverage across reference genome PHE/MN1-00, CDS = Coding sequence

Table 2.8 Comparison of assembly statistics between reference-guided and reference independent assembly (continue)

Sample	Depth of coverage [a]	Assembly method	Genome fraction (%) [b]	N_{50} (bp)	Largest contig (bp)	Total (bp)	Misassemblies	Misassembled contigs length (bp)	No. CDS
3387	14×	Reference-guided	98.5	183,926	344,907	1,694,495	2	487,277	1,424
		Reference independent	98.4	60,837	183,447	1,694,564	5	265,821	1,428
1660	7×	Reference-guided	94.6	4,949	23,430	1,625,190	4	12,562	1,370
		Reference independent	72.3	2,713	18,659	1,245,203	3	26,496	1,060

[a] *L. intracellularis* depth of coverage, [b] Percentage of coverage across reference genome PHE/MN1-00, CDS = Coding sequence

Table 2.9 *L. intracellularis* draft genomes assembled using reference-guided approach with metaSPAdes.

Sample	Sample type	*L. intracellularis* depth of coverage	Genome fraction (%) [a]	N_{50} (bp)	Largest contig (bp)	Total (bp)	Misassemblies	No. CDS
8163	Faecal	3,125×	98.7	391,999	701,716	1,697,347	1	1,418
SRUC1	Faecal	1,594×	98.6	392,018	701,723	1,695,371	1	1,422
661	Faecal	1,081×	98.6	392,020	701,723	1,696,562	1	1,417
5939	Faecal	453×	98.6	392,001	600,122	1,695,680	1	1,422
682	Faecal	158×	98.7	391,994	701,458	1,696,555	1	1,420
976/1	Faecal	147×	98.7	457,371	600,102	1,696,782	1	1,419
2069	Faecal	67×	98.5	335,208	391,928	1,693,281	2	1,438
1886	Faecal	59×	98.7	701,088	727,332	1,697,595	1	1,421
2746	Faecal	55×	98.5	392,026	701,730	1,694,550	2	1,421
630	Faecal	51×	98.7	395,996	701,725	1,698,369	1	1,420
6073	Faecal	30×	98.6	391,969	701,087	1,695,942	1	1,419
5626	Faecal	24×	98.6	392,595	600,104	1,695,140	2	1,422
F22	Faecal	23×	99.6	444,003	727,279	1,713,286	0	1,432
SRUC3	Faecal	19×	99.0	269,950	458,700	1,702,119	1	1,421

[a] Percentage of coverage across reference genome PHE/MN1-00, CDS = Coding sequence

Table 2.9 *L. intracellularis* draft genomes assembled using reference-guided approach with metaSPAdes (continue).

Sample	Sample type	*L. intracellularis* depth of coverage	Genome fraction (%) [a]	N_{50} (bp)	Largest contig (bp)	Total (bp)	Misassemblies	No. CDS
H14	Faecal	17×	98.7	314,686	830,919	1,698,035	1	1,414
H9	Faecal	16×	97.8	217,170	328,180	1,657,068	2	1,416
3387	Faecal	14×	98.5	183,926	344,907	1,694,495	2	1,424
Edinburgh	Colon	34×	98.3	335,336	695,329	1,689,980	1	1,419
Thirsk2	Ileum	21×	98.5	395,982	701,443	1,694,403	1	1,422
4242	Ileum	18×	98.6	335,324	701,737	1,695,089	1	1,417

[a] Percentage of coverage across reference genome PHE/MN1-00, CDS = Coding sequence

2.5 Discussion

In the current study, we have demonstrated that high quality draft genomes of *L. intracellularis* can be acquired through direct sequencing of faecal and tissue samples from pigs and horses positive for *L. intracellularis*. By targeting *aspA*, we were able to accurately quantify *L. intracellularis* DNA from faecal samples, and correlate bacterium copy numbers in the library preparation to sequencing depth of coverage (Figure 2.3A). Our data suggested that for faecal samples, a minimum of 2×10^4 copies of *L. intracellularis* DNA per ng of genomic DNA is required to achieve a sufficient depth of coverage of more than 20×. In the case of ileum, a minimum of 2×10^2 copies of *L. intracellularis* DNA is estimated to achieve sequencing depth of more than 40×.

Generating sufficient depth of coverage for genome assembly through direct sequencing of clinical samples is challenging. This is largely due to the difficulties in obtaining sufficient amount of bacterial DNA, as clinical samples typically contain a high abundance of host genomic DNA and low abundance of bacterial DNA (Methé et al., 2012). Currently, few commercial kits are available to adequately address this problem. The MOlYsis® and QIAamp DNA Microbiome kit rely upon the structural differences between eukaryotic and prokaryotic cell walls to eliminate host DNA by selectively lysing host cells, followed by degradation of the released DNA. SureSelect target enrichment system (Agilent) is a post DNA extraction enrichment method, which uses custom-designed probes as baits to capture target DNA from genomic DNA, followed by sequencing of the captured DNA. Here, we assessed the ability of NEBNext® microbial enrichment kit to increase the sequencing depth of

L. intracellularis from an ileum and faecal sample. This kit utilises differences in CpG methylation abundance between eukaryotes and prokaryotes to separate DNA from the two domains (Feehery et al., 2013). Our results revealed that while enrichment using the NEBNext® kit worked effectively on the ileum sample, leading to a three-fold increase compared to the non-enriched sample (Figure 2.3 and Table 2.4). Enrichment was ineffective on the faecal sample, resulting in a decrease of *L. intracellularis* sequencing depth (Figure 2.4 and Table 2.4). Detection of *L. intracellularis* DNA in the magnetic bead bound fraction for enrichment of the faecal sample suggests non-specific binding of the MBD2 protein, which was not observed for enrichment of the ileum sample (Figure 2.4). This could be due the fact that faecal sample is comprised of a more complex mixture of substances which may interfere with the ability of the MBD-Fc protein to actively bind to densely CpG methylated DNA (Monteiro et al., 1997, Schrader et al., 2012), and to separate eukaryotic from prokaryotic DNA. Furthermore, DNA quality from faecal samples is poor and more fragmented compared to ileum DNA. This may impact upon the binding and interaction of the MBD-Fc protein, which is optimised for long and intact DNA fragments (Feehery et al., 2013). A major disadvantage of the NEBNext® microbial enrichment kit is the low DNA recovery after the enrichment process, which required a high concentration of input DNA to ensure sufficient amount of DNA was recovered for sequencing. This presented a major drawback for enrichment of faecal samples since recovery of genomic DNA from faecal samples is generally low and insufficient for the enrichment process. Thus, we concluded that the enrichment kit is more suited for use with tissue samples.

Currently, there is a lack of consensus on the performance of different metagenomic assembly software (Sczyrba et al., 2017). This is due to the fact that variations in

metagenomic samples, such as the microbial community structure and the complexity of the dataset, as well as the technology used for sequencing, will all influence the performance of different software (Bradnam et al., 2013, Vollmers et al., 2017, Quince et al., 2017). Here, we compared the performance and assembly results from metaSPAde, IDBA-UD and MEGAHIT, on the assembly of four faecal DNA datasets containing varying levels of *L. intracellularis* sequencing depth. Generally, as depth of coverage decreased the assembly became more fragmented (Table 2.5), presumably due to fewer sufficient overlapping of k-mers and less information available to help resolve ambiguous paths. However, there was a lack of consistency in the quality metrics generated by the assemblers for the four datasets. For assembly of sample containing low levels of *L. intracellularis* sequencing depth, MEGAHIT achieved the largest assembly with 87% genome fraction, while metaSPAdes and IDB-UD achieved 72 and 60%, respectively (Table 2.5). This suggests that MEGAHIT is more suited for recovery of genomes presence in low abundance. However, for samples with sequencing depth of ≥14×, metaSPAdes generally produced the most contiguous genome assembly. Furthermore, it was able to generate the longest misassembled contigs length (Table 2.5). This outcome is in accordance with previous results by Quince et al (2017), who which benchmarked metaSPAdes and MEGAHIT on the assembly of gut, ocean and soil sample, and demonstrated that although metaSPAdes generated longer contigs it appeared to be less accurate (Quince et al., 2017). However, it should be considered that the misassembled contigs length is calculated based on the length of contigs containing the misassemblies (Gurevich et al., 2013). Hence, this value will correlate to the contiguity of the genome assembly (indicated by the N_{50} value) and does not necessarily reflect the inaccuracy of metaSPAdes assembly. A special feature of metaSPAdes contributing to its contiguous assembly is the employment of paired DBG (PDBG), which utilise both

read information and preassembled contigs for graph construction at each iteration through small to large k-mer values (Nurk et al., 2017). However, the storage of this information at each iteration is computationally costly and consequently, metaSPAdes required by far the longest runtime and largest memory consumption for assembly, with peak usage of 236 GB (Table 2.6). Even though metaSPAdes produced the best assembly, memory requirement is a limiting factor and using this tool on a desktop computer would not be possible. Evaluation of assembly cost in terms of memory usage, revealed that MEGAHIT achieved the highest efficiency, with almost five times less memory consumption than metaSPAdes (Table 2.6). MEGAHIT implements a "mercy k-mer" approach and a succinct DBG (SDBG) algorithm (a compressed representation of DBG) for assembly, which enables the assembler to cut down on data size during assembly of large and complex datasets (Li et al., 2015). The former approach involves discarding k-mers that appear once during error correction, and rescuing them if they belong to the same read as the erroneous free k-mers and is able to connect these k-mers together within the DBG.

In attempt to improve assembly quality, we applied a reference-guided assembly approach using metaSPAdes, which involves extracting *L. intracellularis* reads from the metagenomic dataset prior to assembly. Our results demonstrated that the reference-guided approach increased assembly contiguity and accuracy, as the number of misassemblies generated decreased (Table 2.8). Furthermore, using this approach, at sequencing depth of 14× we were able to produce assembly with more than 98% genome fraction, lower than the recommended minimum depth sufficient for reference independent genome assembly at 20× (Desai et al., 2013). However, using such an approach may introduce assembly bias towards the reference genome. Thus, we included three reference genomes in the classification database, used for

extraction of *L. intracellularis* reads, in order to reduce bias towards a single reference genome. Furthermore, the Kraken taxonomic classification tool used for extraction of *L. intracellularis* reads utilises a k-mer strategy to classify reads based on mapping of individual k-mer to the lowest common ancestor, rather than an exact sequence match to the reference (Wood and Salzberg, 2014), thereby enabling heterogeneity to be captured. Nonetheless, different isolates from the same species may exhibit considerable genetic variation, especially in accessory regions such as phage or plasmids with dynamic movements within and between species. Therefore, we have to keep in mind that the reliance on reference genome to assist assembly could potentially lead to the absence of strain-dependent genomic regions.

To address this problem, unsupervised binning was applied to assist in the recovery of *L. intracellularis* draft metagenome-assembled genomes (MAG), without the need for a reference. Unsupervised binning using MetaBAT produced *L. intracellularis* genome bins with more than 90% completeness in four samples (Figure 2.5). However, both copies of the 16S, 23S and 5S rRNA gene, plasmid 1 and plasmid 2 were absent in all genome bins. Analysis of contigs statistics revealed that the majority of genomic fragments missing from the bins possessed contigs length of less than 4 kb in length. This is most likely attributed to the fact that the discriminatory power between TNF distance decreases with smaller contig length, making inter-species separation more difficult (Kang et al., 2015). Therefore, the quality of metagenome assembly will have a substantial influence on the quality of genome bins generated. Failure of MetaBAT to place plasmid 1 and 2 into *L. intracellularis* genome bin could also be due to the fact that plasmids are physically separated from chromosomal DNA, may differ in properties such as base-composition and sequence coverage since they replicate extra-chromosomally (Beaulaurier et al., 2018). A difference in these features represents a challenge for binning software to place

plasmids and the chromosomal core genome together in metagenomic samples. Hence, MAGs are often incomplete with extrachromosomal plasmids missing (Parks et al., 2017). To tackle this issue, tools have been applied to assist the recovery of extrachromosomal elements in metagenomic DNA, including Hi-C-based proximity-guided assembly and single-molecule real-time sequencing, which have previously shown promising results in recovering plasmid DNA (Stewart et al., 2018, Beaulaurier et al., 2018). The Hi-C-based proximity-guided assembly approach involves cross-linking DNA molecules that are in close physical proximity within intact cells, followed by sequencing of the ligated DNA, which will enable plasmids and chromosomes originated within the same cellular host environment to be identified (Beitel et al., 2014). Conversely, the single-molecule real-time sequencing approach utilises bacterial DNA methylation signatures for genome binning, as chromosomal and plasmid are thought to share the same methylation profile (Beaulaurier et al., 2018).

The Minimum Information about a Metagenome-Assembled Genome (MIMAG) standards, developed by the Genomic Standards Consortium (GSC) to provide criteria for assessing MAG quality, states that a genome with more than 90% completeness and less than 5% contamination is regarded as a 'high-quality draft' (Bowers et al., 2017). A MAG with completeness of more or equal to 50%, with less than 10% contamination should be classified as 'medium-quality draft', and a MAG that fall below both of these values should be in the category of a 'low-quality drafts' (Bowers et al., 2017). The current gold standard used to measure metagenome assembly quality is based on presence and absence of single copy marker genes. Such an approach has the disadvantage of inaccurately estimating MAG quality, as our results have demonstrated that absence of a 50 kbp genomic region did not impact the level of completeness estimated by checkm (Figure 2.5 and Figure 2.6). Furthermore, single copy marker genes used for MAG quality assessments are often

based on core genes, contamination with accessory genome in the final assembly can therefore be missed and contamination may be underestimated (Parks et al., 2015). Therefore, caution should be applied when analysing the quality of a genome bin using such approach.

Finally, presence of repetitive DNA sequences remains a technical challenge for short read assembly and mapping, which may hinder detection of structural and copy number variations (Treangen and Salzberg, 2012, Periwal and Scaria, 2014). Such variations can confer changes in gene expression and other functional consequences, which may associate with pathogenicity of the bacterial species (Cui et al., 2012, Gaudriault et al., 2008), important for inference of bacterial virulence and pathogenesis. The availability of long reads sequencing technologies such as the MinIon and PacBio may help to resolve these genomic variants, and coupled with short read sequencing a hybrid assembly approach may help to further improve genome assembly.

In the current study, we developed a novel pipeline for obtaining draft genome sequences of *L. intracellularis* through direct sequencing of clinical samples. From a total of 30 metagenomic samples, we were able to recover 20 high quality genome sequences with more than 97% genome fraction, using a reference-guided assembly approach. Although the current method is still limited by the reliance of using reference genomes to assist *L. intracellularis* genome recovery, the ability to obtain genome sequences without a requirement for *in vitro* culture, presents an opportunity to gain new insights into the genome biology of this pathogen.

Chapter 3

Genomic variation and population structure of *L. intracellularis*

3.1 Introduction

L. intracellularis infection is prevalent in pig herds with both subclinical and clinical infections observed (Kroll et al., 2005), but the underlying cause for difference in clinical manifestation is undetermined. Currently, there is a lack of understanding in the evolution and the genetic basis for *L. intracellularis* pathogenesis. To date, only a handful of studies have investigated the genetic diversity of *L. intracellularis* population, mostly by using low resolution molecular typing methods such as variable number tandem repeat (VNTR) and multilocus sequence typing (MLST) (Gebhart et al., 2012, Pusterla et al., 2013). These studies have revealed very limited genetic variation among porcine isolates, but a distinct sequence profile were observed between porcine and equine derived isolates (Gebhart et al., 2012, Pusterla et al., 2013). Genome annotations of isolate PHE/MN1-00, the first complete *L. intracellularis* genome sequenced, revealed many hypothetical proteins with no shared homology to any proteins on the public repository, suggesting this organism may have evolved unique mechanisms of pathogenesis and survival within the pig host (Gyles et al., 2008). Over recent years, rising outbreaks of equine PE in foals have been reported worldwide (Shimizu et al., 2010, Van Den Wollenberg et al., 2011, Pusterla and Gebhart, 2013). Although hyperplasic lesions resulting in weight loss are also observed in horses, clinical signs and pathology differ between horses and pigs (Pusterla and Gebhart, 2009). Previous cross-species experimental infections in pigs and horses have all failed to develop clinical signs, demonstrating that the ability to cause disease is dependent on the species-origin of the isolate (Vannucci et al., 2012b, Sampieri et al., 2013b). In the current study, we performed comparative genomic analysis of 28 isolates to investigate the genetic diversity of *L. intracellularis*,

to identify potential determinants of bacterial pathogenesis, and genomic signatures of host specificity.

3.2 Aims

1. Investigate the genomic variation and phylogenetic relativeness of *L. intracellularis* isolates derived from equine and porcine hosts.

2. Perform comparative analysis to identify the genetic basis for *L. intracellularis* host specificity and pathogenesis.

3. Investigate the time-frame for the evolution of *L. intracellularis*.

3.3 Materials and Methods

3.3.1 Whole genome sequence alignment and phylogenetic inference

Core genome alignment was performed using MAUVE progressive alignment algorithm with default settings (Darling et al., 2010). Locally collinear blocks longer than 200 bp shared among all the genomes were concatenated using the stripSubsetLCBs script implemented in MAUVE. Maximum-likelihood (ML) phylogenetic inference was performed on the core genome alignments using IQTREE v1.6.3 (Nguyen et al., 2014). ModelFinder implemented in the IQTREE software package was used to find the best nucleotide substitution model that fit our dataset (Kalyaanamoorthy et al., 2017). For core genome alignment of the 28 isolates, ModelFinder selected the TVM+F+I model. For core genome alignment of the 25 porcine isolates, ModelFinder selected the HKY+F+I model. ML phylogenetic tree reconstruction was performed using IQTREE (Nguyen et al., 2014), with the best nucleotide substitution model selected by ModelFinder and 1,000 bootstrap replicates.

3.3.2 Variant calling

Sequence reads from the 25 *L. intracellularis* sequenced from this study were mapped to the reference genome PHE/MN1-00 (GenBank GCA_000055945.1) using Snippy v3.0 (Seemann, 2015). As the samples differed in depth of *L. intracellularis* genome coverage, different thresholds were applied accordingly. To avoid false-positive SNP calls, alignment at *tetW* within the prophage genomic island and ribosomal RNA (rRNA) genes were excluded, as these are highly conserved in bacteria leading to nonspecific mapping from other species. Sequence reads were not available in NCBI

for the genome of E40305 (GenBank GCA_001975945.1) and N343 (GenBank GCA_000331715.1) so contigs were shredded into synthetic reads prior to reference mapping. Bam files produced by Snippy v3.0 were visualised using Artemis (Carver et al., 2011). In addition, Harvest v1.2 was used to perform SNP calling to look for concordance with snippy output by which differences in SNPs were identified, manually checked and curated to remove false positives (Treangen et al., 2014). Visualisation of SNP distributions across the genomes of *L. intracellularis* was performed using Ciros (Krzywinski et al., 2009).

3.3.3 Recombination analysis

Recombination analysis was performed using Gubbins v2.3.1 (Croucher et al., 2014) based on whole genome alignments of the 28 *L. intracellularis* isolates aligned using progressiveMAUVE, with the best tree produced by IQTREE as the input tree. Gubbins estimated the relative impact of recombination to mutation on variation along each branch (r/m). A ML phylogenetic tree of the 28 isolates was reconstructed using IQTREE based on SNPs detected outside the recombining regions identified by Gubbins. BratNextGen (Marttinen et al., 2011) was used to confirm results obtained by Gubbins based on whole genome alignments of the 28 *L. intracellularis*. After the proportion of shared ancestry (PSA) tree was constructed, a cutoff of 0.05 was selected which split the tree into 3 clusters. Twenty iterations were used for the recombination learning algorithm, followed by 100 permutation resampling runs on a single processor and a threshold of 0.05 selected for a significance threshold.

3.3.4 Pangenome analysis

The general feature format (gff) files produced by PROKKA were used as input for Roary pangenome analysis with settings "-e -i 95 -s -p 8" (Page et al., 2015). The roary2svg.pl script was used to plot and visualise the Roary output.

3.3.5 Positive selection analysis

Clusters of orthologous groups were identified using the OrthoMCL algorithm (Li et al., 2003) with the following criteria, minimum of 80% coverage in BLAST pairwise alignments, minimum of 85% sequence identity and minimum of 50% length difference. The filtered clusters of orthologous groups were tested for positive selection using the POTION pipeline* (Hongo et al., 2015). Within each cluster, 1-1 orthologues are removed to ensure analysis of single-copy genes. Clusters are then filtered by length with a minimum of 150 bp deviation from the median, minimum pairwise sequence alignment identity of 70% and a minimum of 3 genes per cluster. Trees of individual clusters were constructed with dnaml with 100 bootstrap replicates. Recombination testing was performed using NSS, MaxChi2 and Phi, to remove groups with evidence of recombination. A total of 1,122 orthologous groups remained after the filtering steps, multiple protein sequences within each group were aligned using PRANK and tested for positive selection using codeml- site-model analysis (implemented in PAML), that allows d*N*/d*S* ratio (ω) to vary among sites (Yang, 2007). Using a maximum likelihood approach, codeml tests each aligned orthologue groups for selection models that allows for positive selection (M2 and M8) and compares the likelihood scores with null models for neutral selection (M1 and M7), followed by a

likelihood ratio test comparing the two nested models, M1/M2 and M7/M8, to determine which model most appropriately fits the data.

*POTION analysis ran by Rodrigo Bacigalupe

3.3.6 Dating analysis with BEAST

To estimate TMRC for porcine-associated isolates we used BEAST v2.4.8 (Bouckaert et al., 2014). First, core genomes of 21 porcine isolates were aligned using the progressiveMauve aligner. TempEst was used to assess temporal signal in the data before proceeding with phylogenetic molecular clock analysis (Rambaut et al., 2016). The ML tree constructed using IQTREE was used as input for TempEst, which revealed positive correlation for regression analysis between root-to-tip genetic distance and sampling time. To calculate the statistical significance of regression, the root-to-tip distance calculated by TempEst was imported into R and plotted against isolate sampling date using a simple linear regression model.

Dating analysis was done using BEAST v1.10.1 with the K81 substitution model (A↔C = G↔T, A↔G = C↔T, A↔T = C↔G) and the Strict Clock model (Bouckaert et al., 2014). Three different demographic models (constant, exponential and Bayesian skyline) were explored. Each demographic model was run for 30,000,000 Markov chain Monte Carlo (MCMC) iterations, with sampling every 1,000 iterations. Selection of the best fit demographic model to our data was based on comparison of Bayes factors following each run.

3.4 Results

3.4.1 Population genomic structure of *L. intracellularis* reveals host restricted genetic diversity

To investigate the population structure of *L. intracellularis* we aligned 28 genome sequences, including 21 acquired through direct sequencing of clinical samples, four genomes from cell passaged samples and three genomes retrieved from NCBI (Table 3.1). In total, 25 isolates originated from porcine hosts and three isolates from equine hosts. Alignment of the genomes yielded a core nucleotide sequence alignment of 1,614,292 bp with 11,306 SNPs identified that was used to construct a ML phylogenetic tree and revealed three distinct clades, comprised of a single porcine clade and two equine clades (Figure 3.1A). To infer an accurate phylogenetic reconstruction, Gubbins was employed to remove the effect of potential recombination events. The resulting tree revealed a similar topology to the tree constructed without recombination removed (Figure 3.1B). However, the phylogenetic distances separating the three lineages were reduced. While all of the 23 porcine isolates clustered into a single clade, displaying low sequence diversity with maximum pairwise distance of 228 SNPs. The three horse isolates divided into two distinct lineages, represented by H14 and H9 horse isolates from the UK clustering in one clade and the E40504 strain from the US in the second clade. The two equine lineages were separated from each other by 5,289 core genome SNPs with the E40504 isolate phylogenetically closer to the porcine clade. The UK equine clade displayed greater diversity indicated by the long branch length. An analysis of the average nucleotide identity (ANI) between the H14 equine strain and the PHE/MN1-00 porcine strain revealed that the two shared >99% ANI.

Table 3.1 *L. intracellularis* isolates used for comparative analysis.

Isolate name	Country of origin	Host	Source	Year of isolation	No. CDS	Reference
5189	NA	Porcine	Cell cultured	1993	1,422	
DKp23	Denmark	Porcine	Cell cultured	2003	1,430	
15540	NA	Porcine	Cell cultured	NA	1,418	
Enterisol® ileitis	NA	Porcine	Cell cultured	NA	1,418	
LR189	UK	Porcine	Ileum	1993	1,416	
ED	UK	Porcine	Ileum	2015	1,419	
Thirsk2	UK	Porcine	Ileum	2017	1,422	
630	UK	Porcine	Faecal	2016	1,420	
682	UK	Porcine	Faecal	2016	1,420	
SRUC1	UK	Porcine	Faecal	2016	1,422	
SRUC3	UK	Porcine	Faecal	2016	1,421	
1886	Poland	Porcine	Faecal	2014	1,421	
9761	Poland	Porcine	Faecal	2014	1,419	
661	Poland	Porcine	Faecal	2014	1,417	
5939	Poland	Porcine	Faecal	2014	1,422	
2746	Poland	Porcine	Faecal	2014	1,421	
8163	Poland	Porcine	Faecal	2014	1,418	
6073	Poland	Porcine	Faecal	2014	1,419	
5626	Poland	Porcine	Faecal	2014	1,422	
3387	Poland	Porcine	Faecal	2014	1,424	
2069	Sweden	Porcine	Ileum	2003	1,418	
4242	Sweden	Porcine	Ileum	2003	1,417	
F22	Brazil	Porcine	Faecal	2016	1,432	
PHE/MN1-00	US	Porcine	Cell cultured	NA	1,439	
N343	US	Porcine	Cell cultured	NA	1,434	Sait et al., 2013
E40504	US	Equine	Cell cultured	NA	1,408	Mirajkar et al., 2017
H9	UK	Equine	Faecal	2017	1,416	
H14	UK	Equine	Faecal	2017	1,414	

NA, data not available. CDS, coding sequence

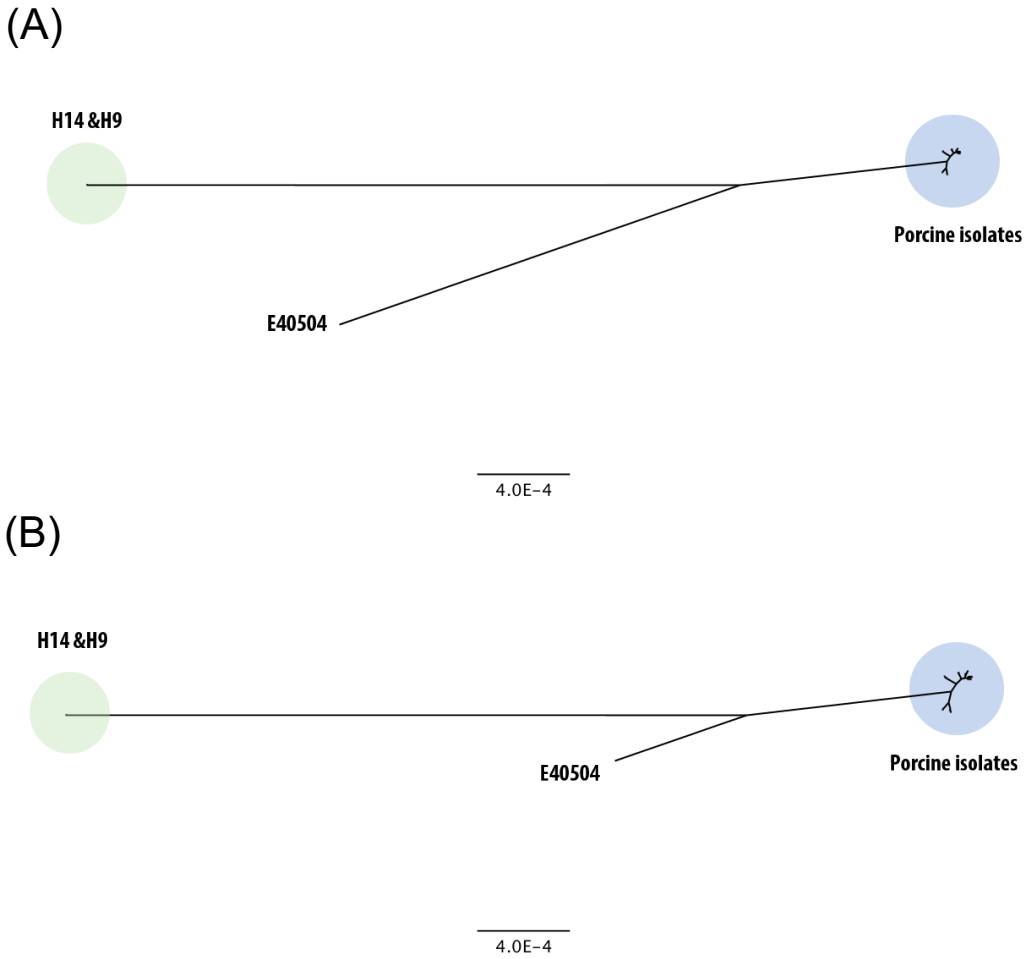


Figure 3.1 Unrooted ML tree generated based on core genome alignments of 28 *L. intracellularis* isolate. (A) ML tree generated without and (B) with recombination regions removed. The ML phylogenetic trees were constructed with IQTREE using the best selected substitution model (TVM+F+I) identified by ModelFinder and bootstrap support value calculated from 1,000 replicates. In both trees **three phylogroups are formed**. The 25 porcine isolates are clustered into a clonal group represented in blue, and the three equine isolates are clustered into two distinct groups, with the two UK equine isolates represented in green. The scale bar represents the number of nucleotide substitutions per site.

3.4.2 Porcine derived *L. intracellularis* isolates revealed limited genetic diversity

To investigate the phylogenetic relationships among porcine isolates, we removed the equine isolates H14, H9 and E40504 and performed comparative genomic analysis of the porcine-derived isolates. The ML phylogenetic tree based on the core genome alignments revealed two major clades (clade I and II) (Figure 3.2). Clade I consists of only European isolates, which can be divided further into a sub-clade with four isolates from the UK and a sub-clade with four isolates from Poland. Isolates in clade II are more geographically diverse, populated by isolates from Europe and the Americas (Figure 3.2). Generally, the phylogenetic distribution of the population is relatively clonal, with an average genetic distance of 149 core genome SNPs separating clade I and II.

Among the 25 porcine isolates we identified a total of 482 polymorphic sites sporadically distributed across the genome (Figure 3.3). Three-hundred seventy-five SNPs were found within CDS and affecting only 21% of the core genes, with the majority of SNPs resulting in amino acid replacements (254 nsSNPs and 121 ssSNPs). Nonsense mutations were identified in five genes: *LI_RS02880*, *LI_RS06380*, *LI_RS06980*, *LI_RS06735* and *LI_RS06900* (PHE/MN1-00 locus tags) (Table 3.2). The gene *LI_RS06980* was predicted to encode a putative membrane-associated protein containing an autotransporter domain. Mutation in this gene was specific to the sub-clade comprised of four isolates from the UK in clade I (SRUC1, ED, 630 and 682). The gene *LI_RS02880* was predicted to encode the flagellar synthesis regulator protein FleN. A mutation in this gene was observed in six monophyletic isolates in clade II (1886, 661, 9761, 8163, 15540 and Enterisol), which resulted in the loss of an amino acid at the terminus of the protein. Finally, a mutation

in *LI_RS06380* (predicted to encode the chemotaxis protein Chew) and *LI_RS06980* (putative autotransporter) was isolate-specific and exclusively identified in the DKp23 and 2069 isolate, respectively.

Pairwise comparison between the Enterisol® Ileitis vaccine strain and isolate 8163, a clinical isolate phylogenetically the closest to Enterisol® Ileitis, identified minor variation of 24 SNPs (13 nsSNPs and 8 sSNPs) (Supplementary Table 1).

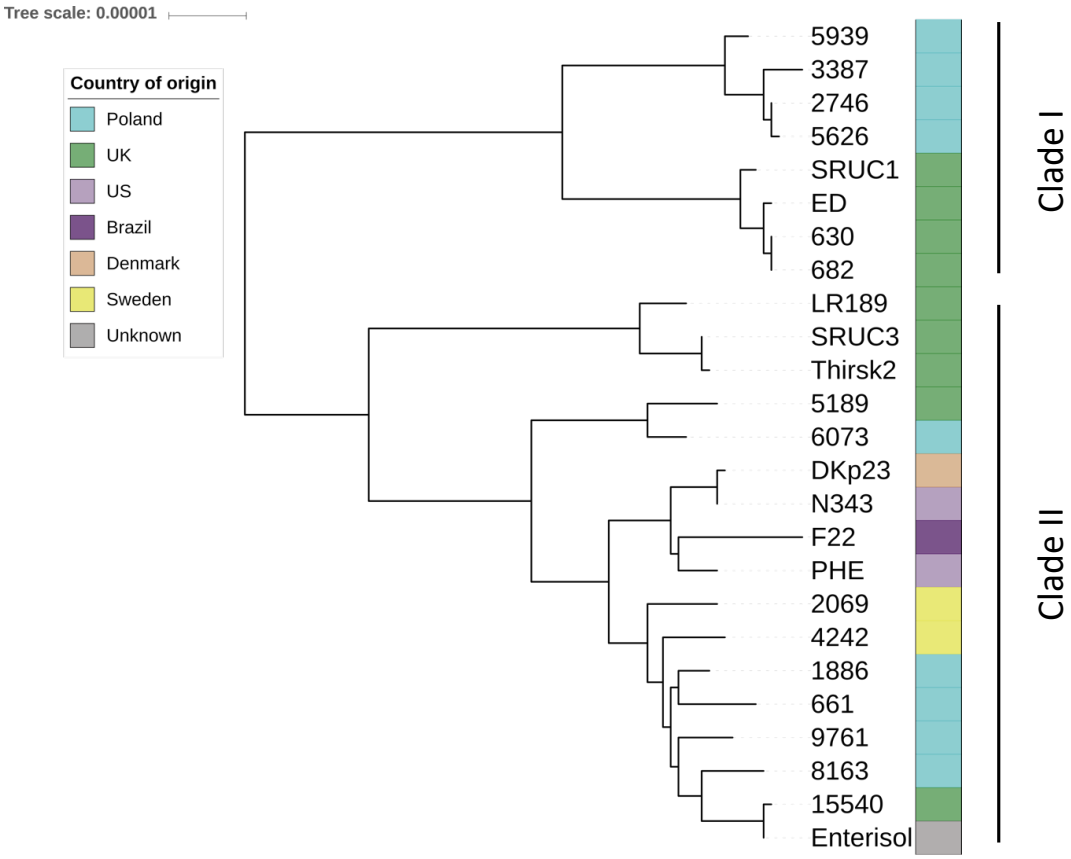


Figure 3.2 ML Phylogeny of porcine *L. intracellularis* isolates across Europe, Central and South America. Two major clades labelled as I and II are identified. The ML phylogenetic tree was constructed based on the core genome alignments of 25 isolates with IQTREE using the best selected substitution model (HKY+F+I) identified by ModelFinder and midpoint rooting. Bootstrap value support was calculated with 1,000 replicates, all nodes displayed contain > 80% support. Different colour blocks along the column of the tree represent the country of isolate origin, when known.

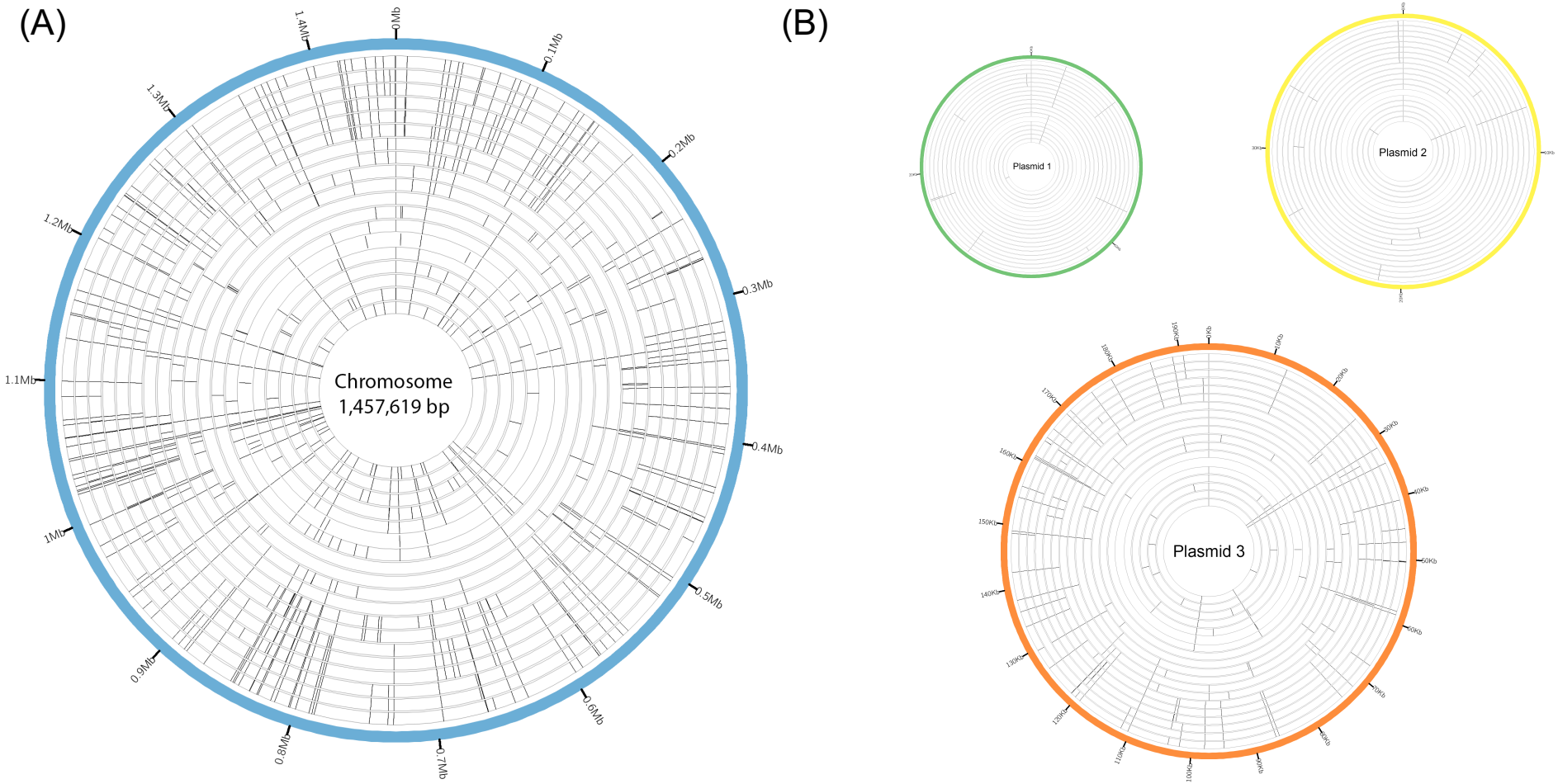

(A)
Chromosome
1,457,619 bp
0Mb
0.1Mb
0.2Mb
0.3Mb
0.4Mb
0.5Mb
0.6Mb
0.7Mb
0.8Mb
0.9Mb
1Mb
1.1Mb
1.2Mb
1.3Mb
1.4Mb
(B)
Plasmid 1
Plasmid 2
Plasmid 3

Figure 3.3 Diagram displaying the distribution of SNPs across the genomes of *L. intracellularis* porcine isolates. Short reads from 19 *L. intracellularis* porcine isolates were mapped against the reference genome of PHE/MN1-00. (A) SNPs identified in the chromosome (B) SNPs identified in plasmid 1, represented by the green ring, plasmid 2 represented by yellow ring and plasmid 3 represented by the orange ring. Redundant isolates with pairwise distance of less than 2 SNPs differences were removed. Individual grey bars within each circle represent a SNP.

Table 3.2 Nonsense mutations identified among porcine associated *L. intracellularis* isolates.

Locus_ID[a]	Ref [b]	1886	661	9761	8163	15540	ENT	SRUC1	ED	630	682	DKp23	2069	2746	3387	5626	5939	Putative protein annotation
LI_RS02880	G	**A**	**A**	**A**	**A**	**A**	**A**	G	G	G	G	G	G	G	G	G	G	flagellar synthesis regulator protein FleN
LI_RS06380	C	C	C	C	C	C	C	C	C	C	C	**T**	C	C	C	C	C	chemotaxis protein Chew
LI_RS06735	A	A	A	A	A	A	A	A	A	A	A	A	A	**T**	**T**	**T**	**T**	Hypothetical protein
LI_RS06900	C	C	C	C	C	C	C	**T**	**T**	**T**	**T**	C	C	C	C	C	C	Hypothetical protein
LI_RS06980	A	A	A	A	A	A	A	A	A	A	A	A	**C**	A	A	A	A	Autotransporter

Nonsense mutations are highlighted in bold, [a] PHE/MN1-00 locus ID, [b] PHE/MN1-00 reference

3.4.3 Pan-genome analysis revealed *L. intracellularis* isolates have highly conserved gene content variation

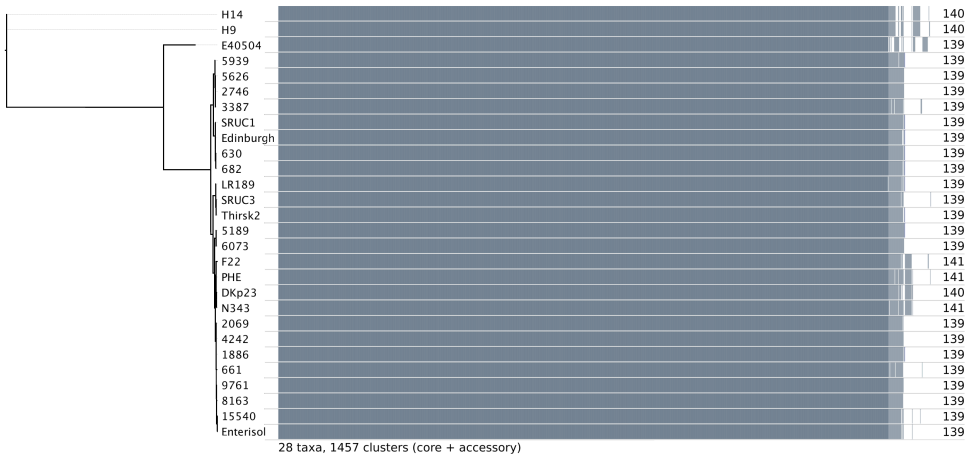
To explore the diversity in gene content variation among *L. intracellularis* isolates we performed pangenome analysis using Roary, which identifies core and accessory genes by clustering homologous groups of genes based on BLASTP sequence identity of derived protein products (Page et al., 2015). A cluster of genes shared by more than 95% of isolates are considered core genes.

Comparative analyses revealed that *L. intracellularis* is highly conserved in terms of gene content, with the total number of predicted genes in each genome ranging from 1,408 to 1,432, a difference of only 24 genes. The total number of unique gene clusters predicted across all 28 isolates was 1,457, with 1,361 (93%) core genes conserved in all 28 genomes (Figure 3.4A). The accessory genome is made up of a total of 96 gene clusters (7% of the pangenome), comprised of 72 partially shared genes common to at least two genomes and 24 strain-specific genes (Figure 3.4B). Fifteen of these accessory genes are located within a previously described prophage-associated genomic island, found only in four porcine isolates, DKp23, F22, N343 and PHE/MN1-00 (Figure 3.4B). These four isolates are monophyletic suggesting that the ancestral prophage was acquired in a progenitor strain. The prophage-associated genomic island is located on the chromosome and possesses a GC content of 60%, higher than the average 33% of the chromosome. The entire length of the genomic island is conserved in the four isolates and within the region encoding a gene for the tetracycline resistance protein TetW (*LI_RS01000*).

A 111 bp in-frame deletion in the gene *LI_RS03480* was observed in three high cell passaged isolates, 15540 (80 passages), DKp23 (23 passages) and Enterisol® Ileitis (live attenuated isolate). While the full-length gene was conserved among the clinical isolates and the low cell passaged isolate 5189 (6 passages). The gene *LI_RS03480* encodes a putative ZIP family metal transporter, belonging to a family of membrane proteins that mediate the uptake of and transport of zinc and other divalent metal cations (Eide, 2005). Closer inspection of *LI_RS03480* revealed the presence of direct repeats flanking the deleted region (Figure 3.5). Intra-molecular recombination between the direct repeats may have resulted in excision of the 111 bp region after a number of cell passages.

Manual inspection of the remaining 81 accessory genes revealed that clusters consisted of divergent gene orthologs due to either protein truncations, in-frame deletions or low amino acid sequence identity (below 95% BLASTP threshold). For example, 23 accessory genes that were identified in the equine isolates (E40504, H14 and H9) initially predicted to be absent from porcine isolates were found to be present but with considerable diversification between 83 to 94 % aa identity (Figure 3.4B).

(A)


H14 140
H9 140
E40504 139
5939 139
5626 139
2746 139
3387 139
SRUC1 139
Edinburgh 139
630 139
682 139
LR189 139
SRUC3 139
Thirsk2 139
5189 139
6073 139
F22 141
PHE 141
DKp23 140
N343 141
2069 139
4242 139
1886 139
661 139
9761 139
8163 139
15540 139
Enterisol 139
28 taxa, 1457 clusters (core + accessory)

(B)

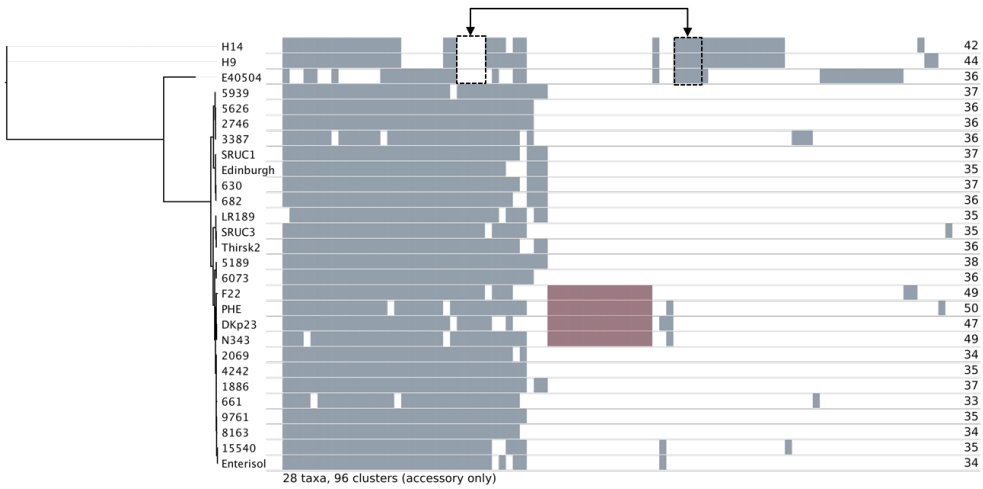

H14 42
H9 44
E40504 36
5939 37
5626 36
2746 36
3387 36
SRUC1 37
Edinburgh 35
630 37
682 36
LR189 35
SRUC3 35
Thirsk2 36
5189 38
6073 36
F22 49
PHE 50
DKp23 47
N343 49
2069 34
4242 35
1886 37
661 33
9761 35
8163 34
15540 35
Enterisol 34
28 taxa, 96 clusters (accessory only)

Figure 3.4 Pan-genome analysis of 28 *L. intracellularis* isolates. (A) ML phylogeny based on core genome SNPs identified outside regions of inferred recombination (left) with core and accessory genes identified by Roary (right). Total number of genes represented in the column on the right. (B) Accessory genes identified in the 28 isolates (left). Aside from the prophage-associated genomic isolate (shaded in red) none of the accessory genes inferred by Roary were acquired through HGT. Instead, these are divergent protein sequences with lower than 95% identity value to the ortholog group. As an example, the outlined boxes represent orthologue genes which are divergent in sequence and placed to the right. Total number of accessory genes represented in the column on the right.

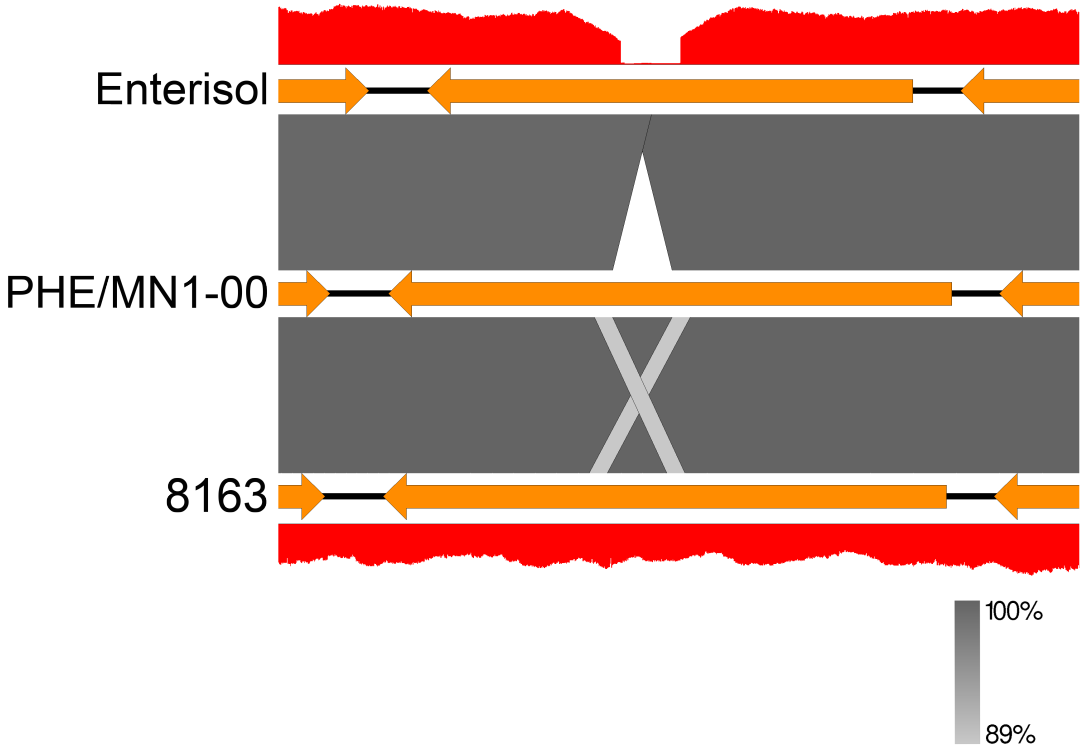


Figure 3.5 Comparison of the putative ZIP family metal transporter locus. Comparison of the gene locus among three isolates. (Top) Enterisol® Ileitis, a live attenuated isolate, (Middle) PHE/MN1-00, the reference genome isolate, and (Bottom) *L. intracellularis* isolate 8163, a clinical isolate phylogenetically closest to the Enterisol® Ileitis strain. The *LI_RS03480* gene, encoding the putative metal transporter is represented by the middle orange arrow. Red coverage graphs at the top and bottom represent sequenced reads coverage at the loci of Enterisol® Ileitis and 8163, respectively. The gradient bar at the bottom right represents percentage of nucleotide BLAST identity. Linear comparison between PHE/MN1-00 and LR189 revealed the presence of repeats flanking the region of deletion in *LI_RS03480*, represented by light grey crisscross bars.

3.4.4 Multiple high SNP density regions across the *L. intracellularis* genome contribute to genetic variation between the porcine and equine clades

Core genome alignments of *L. intracellularis* revealed multiple regions of elevated SNP density across the genome, indicative of recombination. To investigate the impact of recombination on *L. intracellularis* genetic diversity we used two tools, Gubbins and BratNextGen. While Gubbins infer recombination events by searching for regions with high density of polymorphisms relative to the background level (Croucher et al., 2014). BratNextGen employs Bayesian clustering model to detect genetically distinct regions which have been inherited from a separate evolutionary lineage (Marttinen et al., 2011).

Gubbins predicted a total of 43 high SNP density regions along two phylogenetic branches: i) the branch leading to the UK equine clade (consisting of isolate H14 and H9) from the node of diversification (node_31) between the porcine clade and the E40504 isolate, and ii) the terminal branch of E40504 (Figure 3.6). Twenty-seven high SNP density segments were identified on the former branch with a total length of 84,771 bp, accounting for 5% of its core genome (Table 3.3). The impact of these regions on genetic diversification relative to mutation was (r/m) 0.65, indicating diversification along this branch is mainly contributed by mutations (Table 3.3). Sixteen high SNP density regions were predicted along the E40504 terminal branch, with a total length of 165,218 bp affecting 10% of its core genome. The estimated r/m value was 1.8, indicating a greater impact of these inferred recombinant regions over mutations in the genetic diversification of E40504 isolate (Table 3.3). Mapping of these high SNP density regions revealed a random distribution across the chromosome, plasmid 2 and plasmid 3 (Figure 3.7). No recombination events were detected on branches within the porcine clade (Table 3.3).

Using BratNextGen we were unable to detect any significant recombinant regions implying that the putative recombinant events predicted by Gubbins have not occurred between the 28 isolates examined (Data not shown). BLAST search against the NCBI nucleotide database of the high SNP density segments extracted from the H14 genome revealed 89 to 99% identity to the PHE/MN1-00 porcine strain. Due to the small sampling size, especially for the equine isolates, we are unable to determine the origin of these inferred recombinant regions.

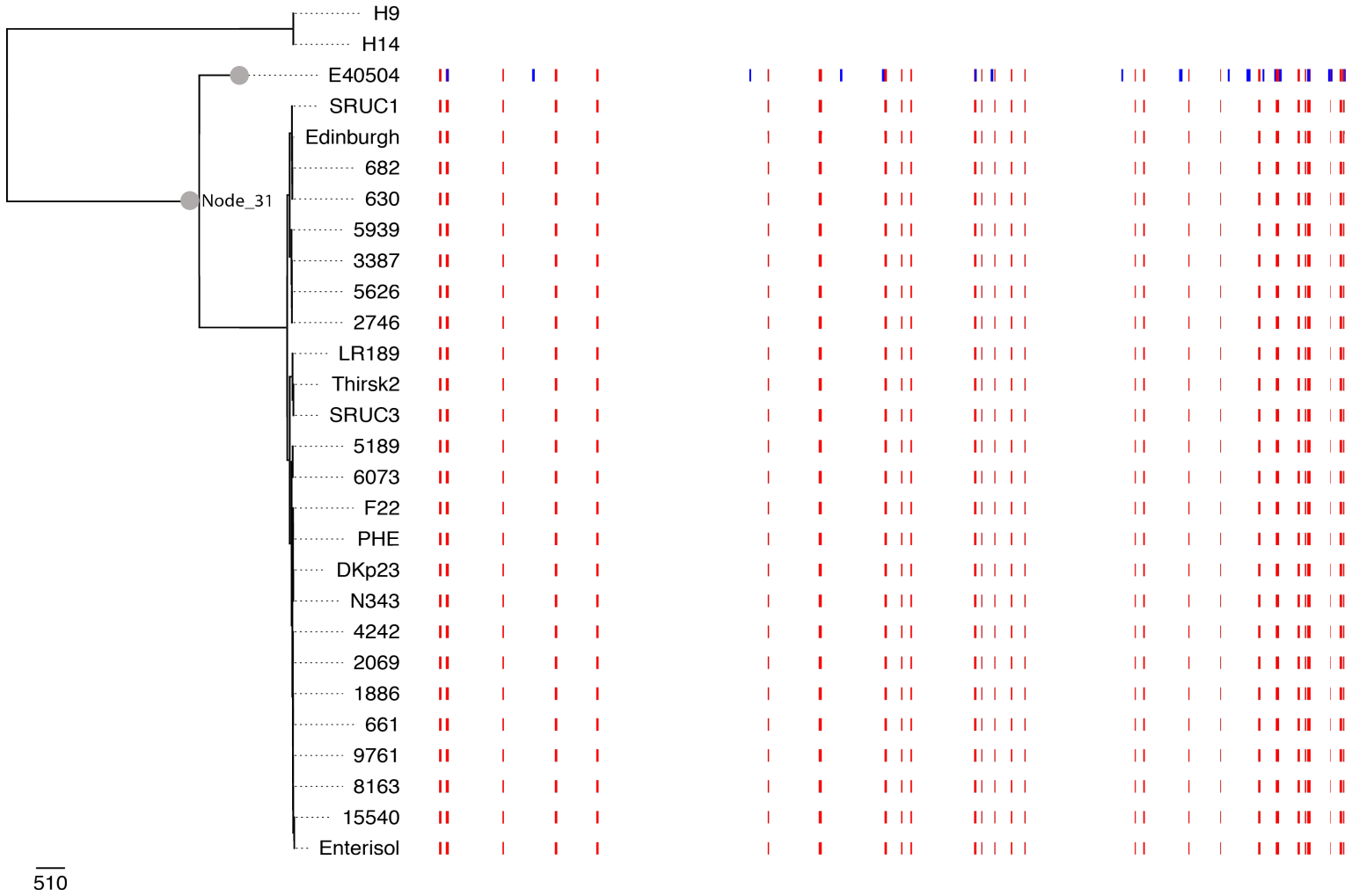

H9
H14
E40504
SRUC1
Edinburgh
682
630
Node_31
5939
3387
5626
2746
LR189
Thirsk2
SRUC3
5189
6073
F22
PHE
DKp23
N343
4242
2069
1886
661
9761
8163
15540
Enterisol
510

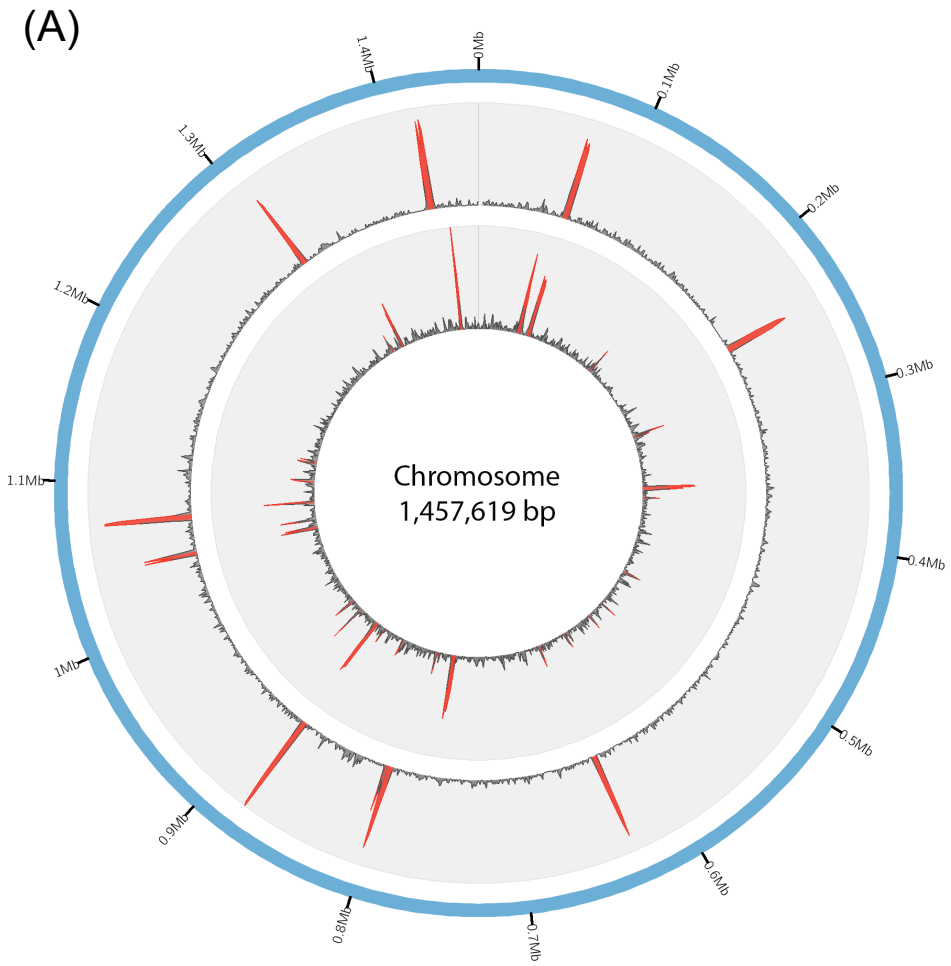
Figure 3.6 Gubbins recombination analysis output. (Left) Phylogenetic tree of 28 *L. intracellularis* isolates constructed using the RAxML GTR model, based on core-genome SNPs corrected for recombination. The phylogenetic tree was out grouped using the H14 isolate. The scale bar represents the number of SNPs. Node_31 is the node of diversification of the UK equine clade from the E40504 isolate and the porcine clade. E40504 and nodel_31 are represented by grey dots. (Right) Regions along the core genome alignment with elevated SNP density detect by Gubbins. Red blocks represent regions identified as putative recombinant regions that are shared with multiple isolates, blue blocks are regions unique to the E40504 isolate.

Table 3.3 Base substitution statistics detected by Gubbins.

Node	Total SNPs	No. of mutation events	No. of SNPs inside recombination	Bases in region of recombination (bp)	r/m[a]	ρ/θ[b]
E40504	4363	1547	2816	146,297	1.82	0.010
node_31	6142	3742	2400	84,771	0.64	0.007

[a] Relative impact of recombination and mutation on the variation accumulated on the branch

[b] Frequency of recombination events relative to point mutations on a branch


(A)
Chromosome
1,457,619 bp
0Mb
0.1Mb
0.2Mb
0.3Mb
0.4Mb
0.5Mb
0.6Mb
0.7Mb
0.8Mb
0.9Mb
1Mb
1.1Mb
1.2Mb
1.3Mb
1.4Mb

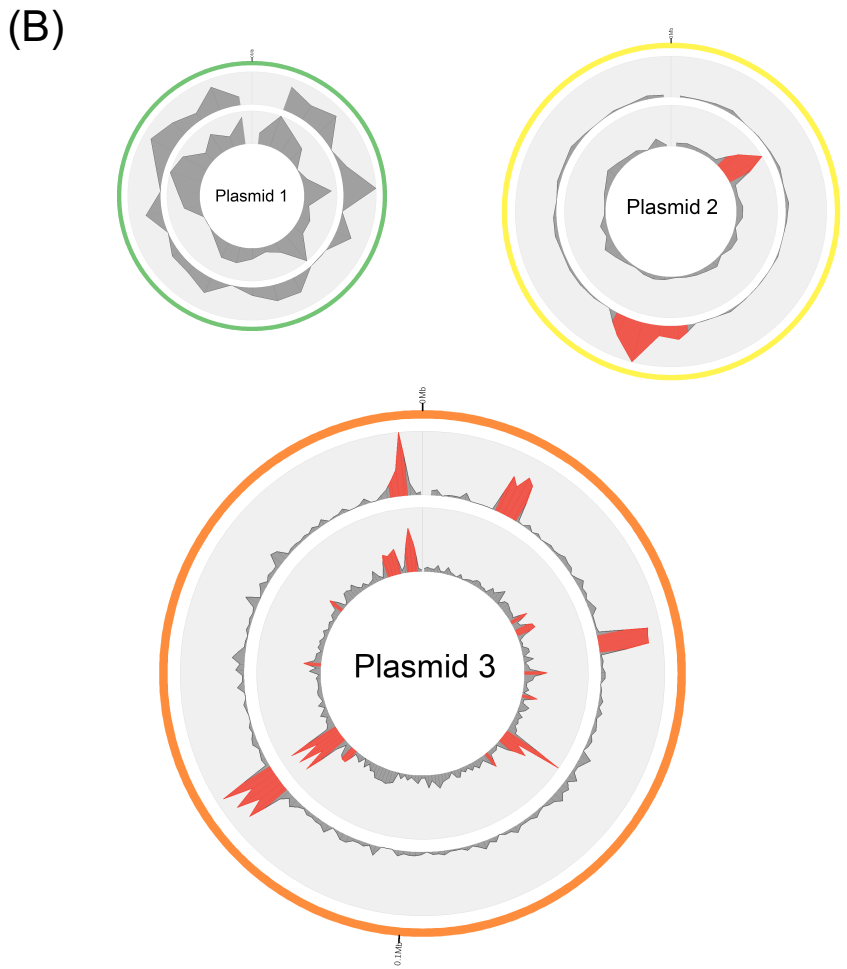

(B)
Plasmid 1
Plasmid 2
Plasmid 3

Figure 3.7 Distribution of regions with elevated SNP density across *L. intracellularis* genomes. SNPs were called by short reads mapping of the UK isolate H14 (outer grey ring) and the US isolate E40504 (inner grey ring) against the porcine isolate PHE/MN1-00 reference genome. Isolates H14 and PHE/MN1-00 were selected as representatives from the two distinct clades. (A) Short reads mapping across the PHE/MN1-00 chromosome, represented by the blue outer ring. (B) Short reads mapping across the PHE/MN1-00 plasmid 1, represented by green ring, plasmid 2 represented by yellow ring and plasmid 3 represented by orange ring. Red bars highlight regions containing more than 10 SNPs per 1kb base region.

3.4.5 Genes involved in energy production, signal transduction and membrane biogenesis are divergent among *L. intracellularis* isolates

A total of 119 genes were identified within the regions of elevated SNP density (Supplementary Table 2). Of which, 19 putative genes encode proteins sharing less than 95% sequence similarity to its ortholog, due to a high abundance of non-synonymous mutations. These are consistent with the gene clusters inferred as accessory genes by Roary from the pangenome analysis. A summary of the putative protein product and its predicted COG function are presented in Table 3.4. Among these, four genes were divergent between the equine and porcine isolates, ten genes were exclusively divergent in the E40504 strain, and five genes were exclusively divergent in strain H14 and H9 (Table 3.4).

BLASTP search of the nine hypothetical proteins against the NCBI database revealed four (LI_RS00755, LI_RS06305, LI_RS06320 and LI_RS07130) were exclusive to *L. intracellularis* with no close homologs identified. Among the remaining five hypothetical proteins, LI_RS06955 shared 46% sequence homology with a lysis protein identified in *Pseudomonas aeruginosa*. The LI_RS06315 predicted to contain a SseC superfamily domain. The gene encoding for this hypothetical protein, together with *LI_RS06310*, which encode for a putative chaperone protein, and *LI_RS06320* belong to a type III secretion operon found to be highly expressed during infection *in vivo* (Vannucci et al., 2013b). Hypothetical protein LI_RS03555 predicted as a putative AT, this protein was previously identified as a prominent antigen from the porcine *L. intracellularis* LR189/5/83 isolate, which was strongly recognised by sera from infected pigs and termed LatA (*L. intracellularis* autotransporter A) (Watson et al., 2011). Finally, the hypothetical protein LI_RS03835, shared closest protein

sequence homology of 50% with a putative glycosyltransferase identified in *Desulfovibrio vulgaris*.

Table 3.4 Genes with highly divergent protein sequences between isolates from equine and porcine clade.

PHE/MN1-00 locus tag[a]	Predicted gene name	Putative protein product	COG
LI_RS00280	*uvrD*	ATP-dependent exoDNAse subunit beta	L
LI_RS07130		Hypothetical protein	
LI_RS07155		Hypothetical protein	
LI_RS07310	*mcpQ*	Methyl-accepting chemotaxis protein McpQ	T
LI_RS03385	*aat*	Leucyl/phenylalanylt-tRNA protein transferase	O
LI_RS03555	*latA*	Hypothetical protein (LatA)	
LI_RS04720	*mviN*	Murein biosynthesis integral membrane protein MurJ	M
LI_RS05825	*fdhF*	Fe-S-cluster-containing hydrogenase	C
LI_RS06305		Hypothetical protein	
LI_RS06315		Hypothetical protein	
LI_RS06320		Hypothetical protein	
LI_RS06710		SAM-dependent methyltransferase	
LI_RS06715		SAM-dependent methyltransferase	
LI_RS06955		Hypothetical protein	
LI_RS00755		Hypothetical protein	
LI_RS01580		Cell wall-associated hydrolase	M
LI_RS03835		Hypothetical protein	
LI_RS03850		Maf-like protein	D
LI_RS06385		5’-nucleotidase	F

Table annotation next page

[a] RefSeq PHE/MN1-00 reference genome locus tag, COG = Clusters of Orthologous Group, yellow shading represents genes which are divergent between equine and porcine isolates, green shading represents genes which are exclusively divergent in the E40504 strain and blue shading represents genes which are exclusively divergent in the H14 and H8 strain. C – energy production and conversion, D – cell cycle control and mitosis, F – nucleotide metabolism and transport, L – replication and repair, M – cell wall/ membrane/ envelop biogenesis, T – signal transduction.

3.4.6 Putative genes involved in host cell invasion and stress response display signatures of positive selection

To look for genes evolving under positive selection among the 28 *L. intracellularis* isolates, we applied the POTION pipeline (Hongo et al., 2015). After sequence filtering, a total of 1,122 orthologous sets containing at least three members per cluster were analysed independently. Codeml identified a total of 11 genes with signatures of positive selection ($P < 0.05$), their putative protein product and predicted COG function are presented in Table 3.5. Among these, the genes *LI_RS02810*, *LI_RS04590*, *LI_RS06305*, *LI_RS00755* and *LI_RS07315* were found within regions containing elevated SNP density, as previously identified.

Several of the genes with signals for positive selection have closest homologues in other enteric pathogens. The gene *LI_RS04590* is predicted to encode for a putative invasin homologous to a porin family protein (sharing 40% protein sequence identity) in *Providencia alcalifaciens*, a Gram-negative bacterium associated with traveller's diarrhoea (Albert et al., 1992). The gene *LI_RS06625* encodes a hypothetical protein with the closest homolog (sharing 55% protein sequence identity) in *Trichuris suis*, a whipworm for which pigs are the natural host. Genes *LI_RS03765*, *LI_RS04845* and *LI_RS02810* all have close homologues in *Bilophila wadsworthia*, sharing 57%, 74% and 52% protein sequence identity, respectively.

Table 3.5 Genes with signatures of positive selection in *L. intracellularis*.

PHE/MN1-00 locus tag[a]	Predicted gene name	Putative protein product	COG	*P* value(M1/M2)	*P* value(M7/M8)
LI_RS06625		Hypothetical protein		0.00736	0.007364
LI_RS00755		Hypothetical protein		0.009973	0.009986
LI_RS04590		Putative invasin	M	0.010852	0.01109
LI_RS00415	*icc*	3',5'-cyclic adenosine monophosphate phosphodiesterase CpdA		0.007291	0.01575
LI_RS06730		Putative molybdopterin cofactor synthesis protein A		0.017054	0.017289
LI_RS07135		Hypothetical protein		0.027472	0.034632
LI_RS03765	*pepA*	Cytosol aminopeptidase	E	0.036495	0.037471
LI_RS06305		Hypothetical protein		0.036030	0.040136
LI_RS03295		Voltage-gated chloride channel family protein	P	0.040088	0.040311
LI_RS04845		Universal stress protein	T	0.043965	0.044044
LI_RS02810	*ispB*	Geranylgeranyl pyrophosphate synthase	H	0.049322	0.049801

[a] RefSeq PHE/MN1-00 reference genome locus tag, E – amino acid metabolism and transport, H – coenzyme metabolism, M – cell wall/ membrane/ envelope/ biogenesis, P – inorganic ion transport and metabolism. T – signal transduction

3.4.7 Estimation of the timeframe for evolution of *L. intracellularis*

Bayesian analysis using BEAST was performed to estimate the evolutionary rate and time of divergence for *L. intracellularis* porcine isolates. The core genomes of 21 porcine isolates sequenced in this study were used for estimating the time of the most recent ancestor (tMRCA). The genome of isolate 15540 was excluded from the analysis as this isolate had undergone 80 cell passages, and may contain mutations that arose from artificial selection during *in vitro* culture, which may interfere with the estimation of its evolutionary rate. In addition, genomes of Enterisol® Ileitis, PHE/MN1-00 and N343 isolates were also removed from the analysis, due to the lack of information regarding the date of isolation and number of cell passages for both strains.

To explore temporal signal within the data set, a regression analysis using TempEst was performed to ensure there are sufficient genetic changes within the time scale of our isolate sampling, so that we can infer a statistical relationship between genetic divergence and time. Statistical significance of the regression line was calculated at p-value of 0.065 (statistically significant value at $p<0.05$). Although not statistically significant, the positive correlations between genetic divergence and isolate sampling time implies sufficient temporal signal is detected within the data set making it suitable for phylogenetic molecular clock analysis (Figure 3.8). Using the three different demographic models, the estimated mean clock rate remained consistent at approximately 3×10^{-7} substitutions per site per year (Table 3.6). However, due to the considerable uncertainty of evolutionary rate for *L. intracellularis* and a lack of prior knowledge, BEAST estimation of mean root age was highly variable for the three different demographic models, with the 95% HPD interval between 76 – 1066 years.

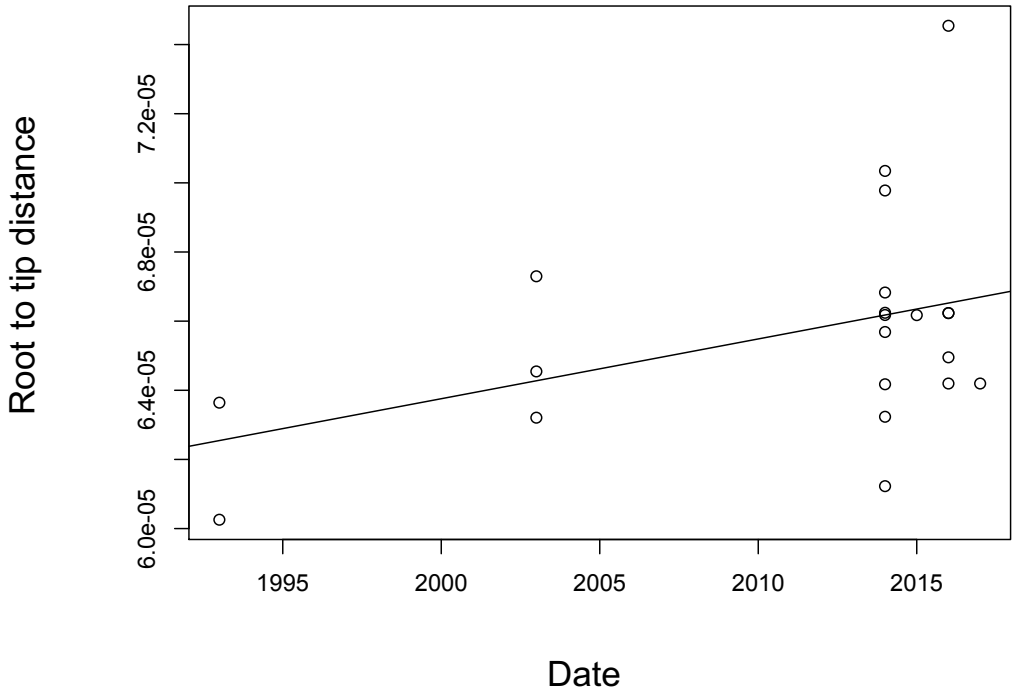


Figure 3.8 Root-to-tip regression panel for the 21 isolates. The root-to-tip genetic distance from a midpoint rooted ML phylogenetic tree of the 21 isolates was measured and plotted against sampling time of each isolate. A positive linear regression line with p-value of 0.065 (statistically significant value at $p<0.05$), although not statistically significant, sufficient temporal signal is detected for molecular clock analysis.

Table 3.6 Estimated evolutionary rate and divergence times for the tMRCA of *L. intracellularis* porcine isolates under different demographic models in Beast2.

Substitution model	Clock model	Demographic model	Mean clock rate	95% HPD interval (Clock rate)	Mean root age	95% HPD interval (root age)
HKY	Strict	Constant	2.915×10^{-7}	$1.07\times10^{-10} – 5.53\times10^{-7}$	630	76 – 1066
HKY	Strict	Exponential	3.616×10^{-7}	$8.18\times10^{-8} – 6.43\times10^{-7}$	232	76 - 483
HKY	Strict	Bayesian Skyline	3.140×10^{-7}	$2.42\times10^{-8} – 5.93\times10^{-7}$	424	76 - 849

HPD = highest posterior density

3.5 Discussion

The current work represents the first study to explore whole genome sequencing of *L. intracellularis* to understand its population genetic structure. A ML phylogenetic tree base on core genome alignment of 28 isolates derived from porcine and equine hosts revealed three distinct clades, comprising of a porcine and two equine clades (Figure 3.1). This is concurrent with previous molecular typing of *L. intracellularis* using low resolution VNTR and MLST approach, which demonstrated a clear distinction between isolates derived from porcine and equine host (Gebhart et al., 2012, Pusterla et al., 2013). Here, we have demonstrated that porcine and equine derived *L. intracellularis* isolates may represent different subtypes of *L. intracellularis* species, as these isolates form phylogenetically distinct host-restricted lineages and sharing genome ANI of more than 99% (Figure 3.1).

The *L. intracellularis* E40504 strain is of equine-origin, originated from the US and have been previously employed in several experimental infections using pigs, horses, rabbits and hamsters (Vannucci et al., 2012c, Sampieri et al., 2013b). The E40504 strain is reported to be host-specific to horses and rabbits, generating serological responses, bacterial shedding and histologic lesions in the animals upon exposure (Vannucci et al., 2012c, Sampieri et al., 2013b). However, the strain is unable to cause infection in pigs and hamsters (Vannucci et al., 2012c, Sampieri et al., 2013b). In this study, equine-derived isolates originated from the UK (H9 and H14) and the E40504 strain form two phylogenetically distant clades, separated by 5,289 core genome SNPs (Figure 3.1). It is worth noting that isolates from the UK are contemporary isolates (from 2017) sequenced directly from faecal samples without *in vitro* culture.

Whereas, the E40504 US isolate had been cultivated *in vitro* for an unknown number of passages and the date of isolation is undetermined (Mirajkar et al., 2017). These variations may contribute to some of the genetic variations observed between the two horse clades, as well as differences in geographical location.

The population structure of *L. intracellularis* suggests the porcine lineage may have emerged from a clonal expansion (Figure 3.1), by which the 25 porcine isolates, sampled from six countries across three continents demonstrated limited genetic diversity, with maximum pairwise distance of 225 SNPs. Phylogenetic analysis of the porcine associated isolates revealed two sub-clades (clade I and clade II). While isolates from the UK and Poland formed phylogeographic clustering across both major clades of the phylogenetic tree (Figure 3.2). Isolates from the US and Brazil were interspersed within isolates from European countries in clade II, indicative of global transmission of *L. intracellularis* from both major clades (Figure 3.2). Sampling of more isolates from different regions within and across countries are required to further enhance this observation.

Pan genome analysis of *L. intracellularis* revealed very limited gene content variation among the 28 isolates, with accessory genes accounting for 7% of its pan genome size (Figure 3.4). *L. intracellularis* display characteristics of a closed pangenome, which is often typical of bacteria with an obligate intracellular lifestyle due to the confinement of its environment with less exposure to exogenous DNA and less opportunity for HGT to occur (Bordenstein and Reznikoff, 2005). The majority of gene content variation was attributed to the presence of a 18kb prophage-associated genomic island identified in four monophyletic porcine isolates (Figure 3.4). This genomic region was previously described by Vannucci *et al* as a DLP12-associated island, which is thought to be defective and have lost its infectious function as genes

encoding for the lysis cassette, normally found in the DLP12 cryptic prophage of *E.coli* K-12, are absent (Vannucci et al., 2013d, Lindsey et al., 1989). It has been previously reported that the presence of this genomic region does not associate with a more virulent phenotype, as PCR analysis of US isolates revealed the presence of this genetic element in all strains, regardless of the clinical presentation (Vannucci et al., 2013d). Indeed, our data agrees with this observation as two isolates (682 and 630) sequenced from pigs that suffered haemorrhagic PE did not harbour this prophage-associated island. From the same study, the authors suggested an association of this genomic region with host-adaptation in pigs since it was observed exclusively in porcine-derived isolates, and absent in isolates derived from clinically affected horses and wild rabbits (Vannucci et al., 2013d). Consistent with this, none of the horse-derived isolates in the current study contain the island. However, we have demonstrated that this region is not required for porcine-adaptation, as it is absent in the 21 porcine-derived isolates examined in the present study. We suspect the presence of the genomic region may be more prevalent in isolates across South and Central America with the two Central America and one Brazil porcine-associated isolates included in this study all harbour this prophage-associated island. Among the 15 genes located within the prophage is *tetW*, encoding for a tetracycline resistance protein. Tetracycline is commonly used in pig production as in-feed antimicrobials to promote growth rate or to prevent or control diseases (Dewey et al., 1999). The prevalence of prophage in *L. intracellularis* isolates observed in the Americas may reflect the higher antibiotic usage in these countries (2017), which have selected for the *tetW* gene. Of note, one isolate originated from Denmark contains the phage-associated island, suggesting that it may also be circulating in European swine. A larger global surveillance study is required to examine the prevalence of the prophage-associated island and thoroughly test this hypothesis.

One genotypic trait which is associated with a high number of cell passaging in our data was a 111 bp in-frame deletion in *LI_RS03480*, encoding a putative ZIP family metal transporter important for maintaining intracellular metal ion homeostasis (Eide, 2005). Family members of this protein have been previously linked to pathogenesis in *S. enterica* serovar Typhimurium to promote bacterial colonisation, and in *E. coli* for growth under zinc-limiting conditions (Cerasi et al., 2014, Grass et al., 2002). This in-frame deletion was one of the few genetic variations identified from the pairwise comparison between the Enterisol® Ileitis strain with a phylogenetically closest clinical isolate (Figure 3.5), and may contribute to the attenuation of Enterisol® Ileitis. Experimental infections of mutant *L. intracellularis* strain with the 111 bp in-frame deletion of the *LI_RS03480* gene should be carried out to test this hypothesis.

A total of 43 high SNP density regions were detected by Gubbins along two phylogenetic branches of the ML tree (Figure 3.6), indicative of recombination events. While putative recombinant regions along the terminal branch of E40504 have a greater contribution towards the genetic diversity of the E40504 isolate, with estimate r/m value of 1.8 (Table 3.3). Divergence of the ancestral branch from the UK equine clade to the node of divergence between the porcine clade and the E40504 horse isolate, appeared to have been mainly the result of point mutations, with an estimated r/m value of 0.67 (Table 3.3). Of note, methods such as Gubbins relying on detection of elevated SNP density to infer recombination may be predisposed to generating false positive inference of recombination. This can occur at loci which are under selection pressure or relaxed constraints, leading to accumulation of mutation at a substantially different rate, to the rest of the genome (Husmeier, 2005). Additional recombination analysis with BratNextGen failed to detect any significant recombination event among the 28 isolates. Due to the small sample size in the current analysis we are unable to infer whether these high SNP density regions are

contributed by recombination from as-yet-unsampled *L. intracellularis*, or may represent regions under selection pressure.

We investigated regions of SNP clustering and identified 19 genes which were polymorphic between *L. intracellularis* isolates among the three phylogenetic clades (Table 3.4). Functional classification of these genes, based on clusters of orthologous groups (COGs) predicted proteins involved in cell metabolism, signal transduction, membrane biogenesis, cell cycle/mitosis control and hypothetical proteins with unknown function (Table 3.4). A region that has a significant excess of SNPs corresponds to a previously reported type III secretion operon. Among which, two genes displayed high sequence diversification between the E40504 strain and porcine isolates, and were among the most highly expressed *L. intracellularis* transcripts in the porcine enterocyte cytoplasm, during peak of infection (Vannucci et al., 2013b). This suggests they may be relevant for host-specific infection of *L. intracellularis* and their exact role during infection warrants further investigation.

To investigate genes with signatures of positive selection among the 28 *L. intracellularis* isolates, we applied likelihood ratio testing of different models for selection which identified 11 genes with p-value of less than 0.05, significant for the positive selection models (Table 3.5). Among these, included a gene encoding a putative invasin, involved in facilitating host cell penetration in many pathogenic bacteria (Cambronne and Schneewind, 2005), and four genes encoding for putative proteins with closest homologues in other enteric bacterial pathogens. Additionally, one gene encoding a hypothetical protein with closest homolog (50% identity) identified in the nematode *Trichuris suis,* in which pigs are the natural host and cause infections with clinical symptoms similar to porcine PE (Beer, 1973). These observations suggests that these genes may have been acquired through HGT from

the host microbiome gene pool. It is worth noting that obligate intracellular bacteria tend to have a higher fixation rate of nonsynonymous mutations due to having a small population size, in which effect of purifying selection is weaker than compared to free-living bacteria (Mamirova et al., 2007). In addition, the relaxed selective constraints in genes associated with metabolic function may lead to accumulation of deleterious mutations (Toft and Andersson, 2010, Moran et al., 2009). Hence, this can inflate the d*N*/d*S* ratio value and potentially overemphasise positive selection. Nonetheless, identification of these genes provides new areas of investigation into putative determinants that may contribute to *L. intracellularis* pathogenesis. There was a lack of evidence for positive selection among the porcine isolates, as majority of the genes identified displayed signature for positive selection between equine and porcine isolates. However, given the limited sample size in our dataset and the clonality of the porcine isolates, there may be insufficient genetic variation to detect diversifying selection.

In the current study, we have shown that the genetic diversity of *L. intracellularis* is host-associated, with isolates derived from equine and porcine hosts belonging to distinct lineages. The porcine-derived isolates are genetically conserved with a lack of diversity and recombination with small variation in gene content, suggesting that these isolates may have emerged through clonal expansion from a single or small number of strains. Furthermore, we found little evidence for adaptive evolution in the genome as shown by the lack of signatures for positive selection. This may suggest that the bacterium is highly adapted to the porcine host. However, given the limited sampling size used in the current study this is not conclusive and sequencing of additional *L. intracellularis* genomes is essential for furthering our understanding of its biology.

Chapter 4

Investigation of a novel *L. intracellularis* surface protein

4.1 Introduction

The live attenuated vaccine, Enterisol® Ileitis was the first commercial vaccine against porcine PE, available since 2001. The immunological effect of the vaccine in protection against *L. intracellularis* infection has been investigated (Riber et al., 2015), but the genetic basis for its attenuation remains unknown. Draft genome sequences of Enterisol® Ileitis, along with a clinical isolate LR189, were obtained in a previous study by Ait-Ali *et al* (unpublished data). Whole genome comparative analysis of Enterisol® Ileitis with a clinical *L. intracellularis* isolate revealed limited genetic variation, but identified a 909 bp in-frame deletion in the gene *LI_RS07080* (PHE/MN1-00 locus ID) in Enterisol® Ileitis (Ait-Ali et al, unpublished data). Located on plasmid 3, *LI_RS07080* encodes a putative autotransporter (AT) and was termed LatB (*Lawsonia* autotransporter B). Previously, comparative genomic analysis revealed length variation in the *latB* gene among 28 clinical *L. intracellularis* isolates.

AT proteins constitute a family of outer membrane-associated proteins belonging to the type V secretion system (Kajava and Steven, 2006). Members of this protein family share common structural motifs, with a signal peptide at the N-terminus responsible for directing the protein across the intracellular membrane. Followed by a passenger domain and a β-barrel translocator domain at the C-terminus, essential for translocation of the passenger domain to the bacterial cell surface. The passenger domain is the functional region of the protein and varies widely in sequence, length and function across family members (Dautin and Bernstein, 2007). Once translocated through the outer membrane, the passenger domain may remain covalently associated with the β-barrel translocator or may be cleaved and secreted (Dautin and Bernstein, 2007). ATs are ubiquitous in Gram-negative bacteria and their function is often linked to bacterial virulence including, adhesion, invasion, biofilm formation and

toxicity (Benz and Schmidt, 2011, Dautin and Bernstein, 2007). These properties make ATs an attractive target for vaccine development against pathogenic Gram-negative bacteria.

In the current study, we explored the potential of LatB as a therapeutic/diagnostic target for *L. intracellularis*. We performed molecular cloning to express recombinant LatB protein for immunological examination and determine if this putative surface protein may have a role in immunological response during *L. intracellularis* infection.

4.2 Aims

1. Determine the variation in the number of passenger domain repeats in *latB* among *L. intracellularis* isolates.

2. Examine the expression of the putative AT during *L. intracellularis* infection.

3. Perform molecular cloning and protein expression to determine the immunological response to the putative AT during *L. intracellularis* infection.

4.3 Materials and Methods

4.3.1 Sequence analysis of LatB

The protein sequence of LatB was obtained from the *L. intracellularis* PHE/MN1-00 strain (NCBI protein accession WP_011527377.1). A sequence similarity search was performed using the BLAST algorithms via BLASTP (Altschul et al., 1997). To predict the signal peptide, the primary sequence was submitted to the SignalP v3.0 server (Bendtsen et al., 2004). Secondary structure prediction was performed using the Phyre2 server (Kelley et al., 2015).

4.3.2 Clinical and cell cultured *L. intracellularis* samples used in the current study

Cell cultured *L. intracellularis* samples were provided by Moredun Research Institute, Edinburgh UK. DNA from Swedish clinical isolates were provided by Magdalena Jacobson and originated from a previous diagnostic study (Jacobson et al., 2004). The DNA was extracted from the distal ileum by phenol/chloroform, precipitated by ethanol and eluted in Milli-Q (Jacobson et al., 2004).

4.3.3 Genomic DNA extraction

Extraction of genomic DNA from Enterisol® Ileitis (Boehringer Ingelheim Vetmedica) and cell cultured isolates was carried out using the DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer's instructions. To concentrate the DNA, the spin column was eluted twice using 50 µl instead of 200 µl of AE buffer. The quality and quantity of DNA were assessed using a Nanodrop™1000 (Thermo Fisher Scientific)

4.3.4 Sequencing of *latB*

All custom oligonucleotides used in this study were designed using DNASTAR Lasergene® Core Suite 9 and purchased from Invitrogen. Primers were designed to target non-repeat regions flanking the passenger domain of *latB*, primer pair sequence: forward 5'-TGTGTATATTTGACAGCTGGAGA-3' and reverse 5'-CAAGGACGACGGCTTATCC-3'.

PCR amplification of the *latB* passenger domain from *L. intracellularis* was performed using 10 µM forward and reverse primers, approximately 50 ng of template DNA, 10 mM dNTPs (Roche), 2 U FastStart Taq® DNA polymerase (Roche), PCR reaction buffer 10x concentrated with 20 mM $MgCl_2$ (Roche), GC-rich solution 5x concentrated (Roche) and nuclease free H_2O to a final volume of 25 µl. The thermocycler programme included: an initial denaturation at 95 °C for 5 min, followed by 35 cycles of denaturation at 95 °C for 45 sec, annealing at 58 °C for 45 sec, extension at 72 °C for 60 sec, followed by final extension at 72 °C for 10 min.

PCR amplified products were resolved on 1 % (w/v) agarose gels containing agarose (Invitrogen) suspended in 1 x Tris-acetate-EDTA (TAE) buffer and 1 x SYBR Safe DNA gel stain (Invitrogen). Samples were loaded with 5 x DNA loading buffer blue (Bioline) and electrophoresed with either 1 kb hyperladder (Bioline) or Quick-Load® Purple 2-Log DNA ladder (NEB). Electrophoresis was typically conducted at 110 V for 45–120 min followed by gel visualisation using a G:box (Syngene).

QIAquick® PCR purification or gel extraction kit (Qiagen) was used to purify PCR products from agarose gels following the manufacture's guide. The quality and

quantity of PCR amplified DNA were assessed using a Nanodrop™1000 (Thermo Scientific).

Sanger sequencing was performed by Edinburgh Genomics, University of Edinburgh. FinchTV 1.4.0 (Geospiza) was used to assess the quality of the chromatogram data and analysis was carried out using SeqMan of DNASTAR® Lasergene® 12.2 software.

4.3.5 Synthetic LatB peptide and anti-LatB IgG antibody

The synthetic peptide of a passenger domain long repeat unit was purchased from Perbio Science UK Ltd peptide sequence: LSSESSFDDGHNPSPP (16aa). Polyclonal rabbit anti-LatB IgG antibody generated against this synthetic peptide was purchased from Perbio Science UK Ltd.

4.3.6 Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis and Coomassie Blue Staining

Protein samples were suspended in 5 x SDS loading dye (0.5 M DTT, 50 % (v/v) glycerol, 1M Tris-HCL pH 6.5, 0.25 % (w/v) bromophenole blue and 10% (w/v) SDS) and boiled at 95^{o}C for 5 min. 10 µl of protein samples was applied to 4-20% Mini-PROTEAN®TGX Stain-Free™ Precast Gels (Bio-Rad). Samples were loaded alongside EZ™ Prestained *REC* protein ladder (Fischer Scientific). Electrophoresis was carried out at 150 V for 60 min using Mini Trans-Blot® Cell (Bio-Rad). Proteins resolved on SDS-PAGE were visualised by staining in Coomassie Brilliant Blue High Gel Staining Solution and imaged with an Epson scanner.

4.3.7 Western blot analysis

Proteins resolved by SDS-PAGE were transferred onto nitrocellulose membranes (Bio-Rad) using a tank transfer method with thick blot filter paper (Bio-Rad) at 120 V for 60 min in Transfer Buffer (20 mM Tris Base (Fisher Scientific), 154 mM glycine (Sigma-Aldrich), and 20 % (v/v) methanol (Fisher Scientific) made up to 1 litre with dH_2O), chilled on ice. Membranes were blocked for 1 h at room temperature in 5 % (w/v) skimmed milk powder (Sigma-Aldrich) in PBST (PBS, 0.1 % Tween-20). The primary antibody was applied to the blocking buffer and incubated overnight at 4 ^{o}C. Membranes were washed with PBS for 3 x 15 min, before incubation with secondary antibody in 5 % (w/v) skimmed milk powder (Sigma-Aldrich) in PBST (PBS, 0.1 % Tween-20) for 1 h at room temperature. After another wash in PBS for 3 x 15 min, blots were visualised using 1:1 mix of Pierce™ ECL blotting substrate solution (Thermo Fischer Scientific) and exposed using Amersham Hyperfilm ECL (GE Healthcare Life Sciences).

To confirm the specificity of rabbit anti-LatB IgG for LatB in *L. intracellularis*, Enterisol Ileitis® (Boehringer Ingelheim Vetmedica) was re-suspended in 20 ml PBS and centrifuged at 150 x g for 10 min to pellet the McCoy cells. The cell pellet was washed in PBS and suspended in 5 x SDS loading dye, boiled at 95 ^{o}C for 5 min and resolved on SDS-PAGE, as previously described. Western analysis was performed using rabbit anti-LatB IgG, at 1:1000 dilution as primary antibody and goat anti-rabbit conjugated HRP secondary antibody (Vector Labs), at a 1: 500 dilution as secondary antibody.

4.3.8 Immunohistochemistry

Ileum tissue sections were obtained from a previous challenge study by McIntyre *et al* (MacIntyre et al., 2003a). The pigs were tested negative for *Brachyspira hyodysenteriae*, *B. pilosicoli, Yersinia* spp. and *Salmonella* spp and were PCR negative for *L. intracellularis*. Pigs were challenged with a pure culture of *L. intracellularis* (isolate LR189/5/83) and euthanized at 3, 7, 14, 21, 28, 35 or 42 post challenge (MacIntyre et al., 2003b). Ileum sections were paraffin-embedded and prepared by the pathology department at the Royal (Dick) School of Veterinary Studies. Embedded sections were dewaxed by serial washing of the glass slides with xylene for 3 x 2 min, 99 % ethanol for 3 x 2 min and distilled water for 2 min. Antigen retrieval was performed by digesting the sections with proteinase K (Dako) for 10 min at room temperature. After serial washes in PBS for 10 min, endogenous peroxidase activity was blocked using 3 % (v/v) H_2O_2 for 30 min. Sections were blocked with NGS/PBS/BSA 5 % for 30 min, then co-stained with anti-LatB and anti-VPM53 (originated from previous project) (MacIntyre et al., 2003b), both at 1: 400 dilution in blocking solution, and incubated overnight at 4 ºC. Slides were serially washed in PBS, then stained with goat anti-rabbit IgG Alexa Flour® 647 (Life Technologies) and goat anti-mouse IgG FITC (Life Technologies), both at 1: 1000 dilution in blocking solution for 1 h at room temperature. After another round of serial washing in PBS, slides were stained with DAPI (Sigma Aldrich), diluted at 1:500 in PBS for 10 min at room temperature. Following the final serial wash with PBS, cover slips were mounted onto the slides with Lab Vision™ PermaFlour™ aqueous mounting medium (Thermo Fisher Scientific). Sections were viewed using an LSM 700 confocal laser scanning microscope (Carl Zeiss).

4.3.9 *latB* cloning for recombinant protein expression

Primers used in the current study are listed in Table 4.1, designed and purchased as previously described. High-fidelity PCR was used to amplify *latB* from Enterisol® Ileitis and 4242 genomic DNA using 10 μM forward and reverse primers, approximately 100 ng of template DNA, 10 mM dNTPs (Promega), 1 x *PfuUltra*™ II reaction buffer (Stratagene), 1 U *PfuUltra*™ II Fusion HS Polymerase (Stratagene) and dH_2O to a final volume of 50 μl. The thermocycler programme included: an initial denaturation at 95 °C for 2 min, followed by 30 cycles of denaturation at 95 °C for 20 sec, annealing at 52 °C for 20 sec, extension at 72 °C for 1.15 min, followed by final extension at 72°C for 3 min.

PCR amplified products and pGEX6P1 vector were digested with EcoRI (40 U/ μl) and NotI (40 U/ μl) restriction enzymes (New England BioLabs) in 20 μl reactions for 2 hr at 37 °C. Digested products were resolved by 1 % (w/v) agarose gel, the bands of correct size were excised from the gel and DNA extracted using the QIAquick Gel Extraction Kit (Qiagen) following the manufacturer's instruction. Ligation reactions were carried out using T4 DNA ligase (400 000 U/ μl) (New England BioLabs) in a 15 μl reaction for 1 h at room temperature: reaction mixtures were formulated depending on a 3:1 molar ratio of insert: plasmid.

Plasmid was transformed into *E. coli* XL1-Blue competent cells (Invitrogen) via heat shock-transformation. Competent cells were thawed on ice for 5 min and 10 ng of plasmid was added to 20 μl of cells with 1 μl β-mercaptoethanol (Sigma-Aldrich). The mixture was chilled on ice for 20 min then heated in a shaking water bath at 42 °C for 30 sec, immediately followed by chilling on ice for 5 min. 1 ml super optimal broth (SOB) was added to the cell mixture and placed in a shaker at 180 rpm, at 37 °C for

45 min. The transformation mixture was centrifuged for 1 min and the pellet suspended in 5 µl PBS and plated onto a LB agar plate supplemented with ampicillin (LB/Amp), and incubated at 37 ºC overnight.

Colonies were screened for ligated plasmid by colony PCR, re-plated onto new LB/Amp plates and incubated overnight at 37 ºC. Colony PCR was performed using FastStart Taq® DNA polymerase (Roche), PCR reaction mix and the thermocycler programme described in section 4.3.2. Positive transformed colonies were sub-cultured in fresh LB medium containing ampicillin, and placed in a shaker at 200 rpm overnight at 37 ºC. Ligated plasmids were extracted from transformed cells using the QIAquick® Spin MiniPrep kit (Qiagen) following the manufactures guide. Plasmid DNA quality and quantity were assessed using a Nanodrop™1000 (Thermo Scientific). Insertion of cloned sequence in plasmid DNA was verified by Sanger sequencing described in section 4.3.4.

Table 4.1 Primers used for *latB* cloning for recombinant protein expression.

Primer name	Sequence (5' to 3')*	Function
LIC*EcoR1* F	CCGGAATTCACAGCTGGAGAAGCCTTCTCTATTTC	Cloning of: rLatB ENT and rLatB ENT::Rep, rLatB 4242 and rLatB 4242::Rep
LICRep*Not1* R	ATGCGGCCGCACAAGGACGACGGCTTATCCCATC	Cloning of: rLatB ENT::Rep, rLatB 4242::Rep
LIC*Not1* R	ATGCGGCCGCATTTGAACAGTAACTATCTAATG	Cloning of: rLatB ENT::Rep, rLatB 4242::Rep
pGEX6P1 F	ATCCTGACTTCATGTTGTATGACGC	Integration PCR, Sanger Sequencing
pGEX6P1 3R	ATCCTGACTTCATGTTGTATGACGC	Integration PCR, Sanger Sequencing

*Underlined sequences represent restriction digestion site

4.3.10 Recombinant Protein Expression

Recombinant protein expression plasmids were stored in *E.coli* DH5α cells (Agilent Technologies) at -80 °C and transformed into *E.coli* BL21(DE3)/pLysS (Promega) prior to induction, as described in section (4.3.9). Freshly transformed cells were plated onto LB/Amp agar plates and sub-cultured in LB broth supplemented with ampicillin and placed in a shaker at 200 rpm overnight at 37 °C. 100 µl of sub-cultured cells were used to inoculate 10 ml of fresh LB broth supplemented with ampicillin and incubated at 37°C with shaking at 180 rpm until an OD_{600} of 0.6 to 0.8. When the desired cell density was reached, 1 mM isopropyl β-D-1thiogalactopyranoside (IPTG) (Sigma-Aldrich) was added, cells were incubated at 30 °C with shaking at 180 rpm and induced for 2 hr. 1 ml of induced and un-induced samples were analysed for protein expression by SDS-PAGE, as described in section (4.3.6). The remainder of the cell culture was pelleted at 5000 rpm for 20 min and the pellet stored at -20 °C before progressing for purification.

4.3.11 Native purification of recombinant Glutathione S-transferase (GST)-tagged proteins

Cell pellets were thawed on ice before lysis, followed by suspension in 6 ml lysis buffer (140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, PBS, pH 7.4) supplemented with cOmplete protease inhibitors (Roche Life Sciences). Bacterial lysis was carried out using the OneShot (constant Systems) with a single pulse at 25 kpsi. Cell lysates were centrifuged at 13000 rpm for 20 min at 4 °C and the supernatant passed through 0.45 µm filters (Millipore). Native purification was carried out using a peristaltic pump P-1 (GE Healthcare) and 5 ml pre-packed GSTrap™ 4B column (GE Healthcare). The

column was equilibrated by washing with binding buffer (140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, PBS, pH 7.4) before the lysate was applied. This was followed by further washing of the column with binding buffer before the protein was eluted from the column using elution buffer (50 mM Tris-HCl, 20 mM reduced glutathione, pH 8.0). Fractions were analysed by SDS-PAGE. Quantification of protein was performed with a Nanodrop™1000 (Thermo Scientific) and aliquots were stored at -20 °C.

Western blot analysis of purified recombinant protein was carried out as described in section (4.3.7). The GST-tag was detected using anti-GST tag antibody clone DG122-2A7 HRP conjugate (Millipore) at 1:1000 dilution in blocking solution.

4.3.12 Preparation of *L. intracellularis* Enterisol® Ileitis surface antigens

L. intracellularis surface antigen extraction was prepared by suspending Enterisol® Ileitis (Boehringer Ingelheim Vetmedica) in 20 ml PBS and pelleting at 150 x g for 10 min to separate the McCoy cells from the supernatant containing whole cell *L. intracellularis*. Supernatant was centrifuged at 5000 x g for 20 min and the pelleted bacteria were washed for 4 x in T-NaCl (10 mM Tris-HCl, 50 mM NaCl, pH 8.0) and centrifuged at 8000 x g for 20 min. Followed by suspension in T-NaCl with 0.1 % sodium deoxycholate (DOC) and incubation at 37 °C for 15 min, bacteria were washed twice in T-NaCl and centrifuged at 23,7000 x g for 20 min. Bacteria were suspended in T-NaCl with 0.5 % DOC and incubated at 37 °C for 15 min to release surface antigens. Finally, bacteria debris were pelleted by centrifugation for 15 min at 23,000 x g, the remaining supernatant contained the surface antigens. Quantification of protein was performed with a Nanodrop™1000 (Thermo Scientific) and aliquots were stored at -20 °C.

4.3.13 Enzyme-Linked Immunosorbent Assay

96-well MaxiSorp® plates (Nunc) were coated with GST-tagged rLatB recombinant proteins or *L. intracellularis* surface antigens diluted in coating buffer (15 mM Na_2CO_3, 35 mM $NaHCO_3$, 3 mM NaN_3, pH 9.6) to 1 $\mu g ml^{-1}$, at 4 ^{o}C overnight. Plates were washed with PBST (PBS, 0.1 % Tween-20) for 5 x before blocking with 5 % (w/v) skimmed milk powder (Sigma-Aldrich) in PBST for a minimum of 1 hr at room temperature. Following blocking, plates were incubated with 10-fold serial dilutions of pig sera in PBST for 1 h at room temperature. Plates were washed 5 x with PBST and coated with HRP-conjugated goat anti-swine IgG (Jackson Immuno Research) at 1:10,000 dilution in PBST for 1 h at room temperature in the dark. After washing for 5 x in PBST the plate was analysed using 50 $\mu g well^{-1}$ of ultra-tetramethylbenzidine (TMB) for 2 min at room temperature before the reaction was stopped by addition of 50 µg/well of 2 M sulphuric acid per well. The plates were analysed using a Synergy™HT plate reader (BioTek) at 450 nm wavelength. Data was analysed and presented using Prism 6 (GraphPad). Multiple comparisons were performed when appropriate with one-way ANOVA analysis.

Pig serum samples were kindly by Eleanor Watson from Moredun Research Institute.

4.4 Results

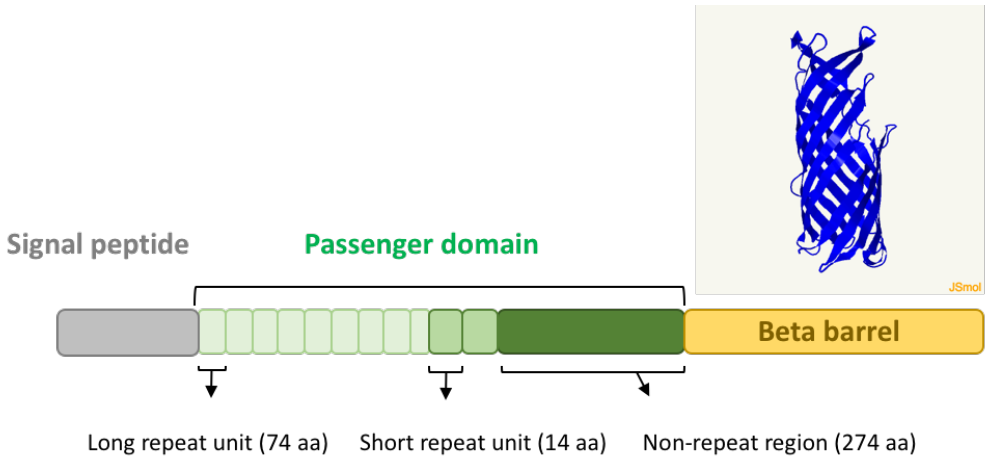
4.4.1 Structural prediction of LatB

A BLASTP analysis of the LatB protein sequence of PHE/MN1-00 strain using the NCBI database was performed to predict conserved domains that could infer a putative function for LatB. The analysis predicted a conserved domain at the C-terminal, 150 aa in length and encoding a AT superfamily β-barrel domain (residues 825 – 975). With the exception of this domain no other conserved motif or domains were predicted. The BLASTP similarity search revealed homologues only in *L. intracellularis* isolates with all homologous proteins predicted as putative ATs. To determine if LatB may share structural homology with experimentally determined proteins, structural modelling of LatB was performed using the Phyre2 Server to predict secondary structure using a template-based homology modelling approach (Kelley et al., 2015). Phyre2 prediction was only able to model 26% of residues with more than 90% confidence. The β-barrel translocator domain was modelled based on the EstA protein (Autotransporter Esterase) of *Pseudomonas aeruginosa* with 99.6% confidence but only 14% sequence identity (Figure 4.1A) (van den Berg, 2010).

A signal peptide at the N-terminal of LatB was predicted using the SignalP 3.0 Server, with a cleavage site predicted at 23 aa (Figure 4.1B) (Petersen et al., 2011). Manual inspection of the LatB sequence between the predicted N-terminal signal peptide and the C-terminal β-barrel translocator domain revealed the presence of 8.5 tandem repeats comprised of 47 aa per repeat unit (termed long repeat units, LRUs), followed by 2 tandem repeats comprised of 14 aa per repeat unit (termed short repeat units,

SRUs) (Figure 4.1A). Adjacent to the repeating units is a 274 aa stretch of non-repeat region.

(A)


Signal peptide
Passenger domain
Beta barrel
JSmol
Long repeat unit (74 aa)
Short repeat unit (14 aa)
Non-repeat region (274 aa)

(B)

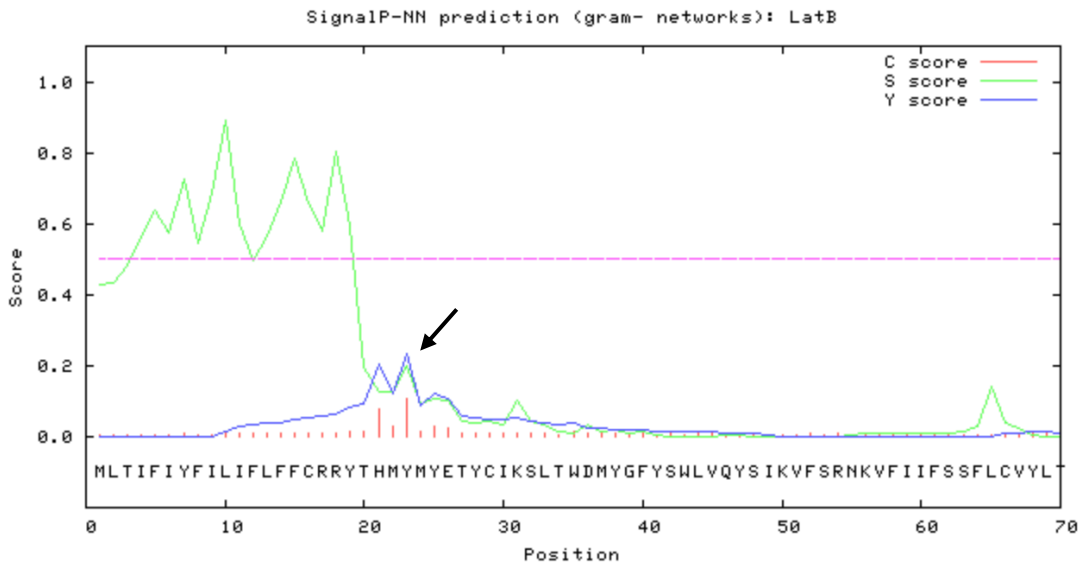

SignalP-NN prediction (gram- networks): LatB
C score
S score
Y score
Score
1.0
0.8
0.6
0.4
0.2
0.0
MLTIFIYFILIFLFFCRRYTHMYMYETYCIKSLTWDMYGFYSWLVQYSIKVFSRNKVFIIFSSFLCVYL
0
10
20
30
40
50
60
70
Position

Figure 4.1 Schematic diagram showing the domain structure of LatB from PHE/MN1-00 strain. (A) The signal peptide (SP) at the N-terminal, followed by the passenger domain comprised of 8.5 long repeat units, 2 short repeat units and a non-repeat region of 274 aa residues. The C-terminal (residues 790 – 1050) contains a β-barrel translocator domain. The secondary structure of the translocator domain was predicted using the Phyre2 server and modelled based on the template of EstA of *P. aeruginosa* (pdb 3KVN) with 99.6% confidence, 14% sequence identity. (B) Screenshot of the SignalP 3.0 server prediction output with a cleavage site predicted at the 23 aa residue. The C-score, representing the raw cleavage site score, should be high at the position immediately after the cleavage site. The S-score, representing the peptide score, should distinguish positions within the signal peptide from positions in the protein without the signal peptide. The Y-score, representing the combined cleavage site score, predicts the cleavage site based on the C-score and the slope of S-score. Based on the Y-score, a cleavage site is predicted at residue position 23, as indicated by arrow.

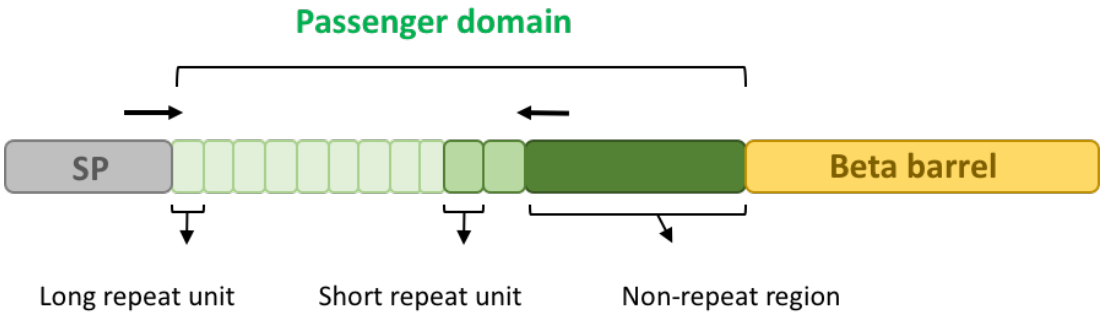
4.4.2 Variation in the length of *latB* were observed among *L. intracellularis* isolates

To investigate the extent of variation in the LatB passenger domain repeat between *L. intracellularis* clinical isolates, PCR analysis and Sanger sequencing was employed in 10 clinical samples obtained from pigs with PE and the Enterisol® Ileitis strain. Comparative analysis showed variation of *latB* was limited to the number of long and short repeat units among the isolates, all of which were in-frame. While the number of long repeat units ranged between 3.5 and 8.5 repeats, the number of short repeat units ranged between one to two repeats (Figure 4.2B and Table 4.2). Variation in *latB* was observed among isolates derived from the same herd (Table 4.2). The shortest variant was 204 aa in length identified in the Enterisol® Ileitis strain, comprised of 3.5 long repeats and 2 short repeats. The longest variant was 439 aa in length identified in the 4242 clinical isolate, comprised of 8.5 long repeats and 2 short repeats (Table 4.2).

To determine if cell passaging may result in loss of repeat units, we assessed length variation in *latB* from cell passaged *L. intracellularis* isolates. From an archive of cell cultured *L. intracellularis*, four strains that have undergone various numbers of cell passages were selected. In general, we did not see a correlation between the number of passages and reduction in the number of LatB passenger repeat length, as Sanger sequencing results revealed the numbers of repeating units remained conserved in all the isolates passaged between 6 to 28 times, with the exception of isolates LR189/02 in which two long repeat domains were lost between passage 12 to 23 (Table 4.3).

Overall, the number of long and short repeat units in the *latB* passenger domain in *L. intracellularis* are highly variable. Our data indicates that *latB* variation is strain dependent, and cell passaging can result in the loss of long repeat units.

(A)


Passenger domain
SP
Beta barrel
Long repeat unit
Short repeat unit
Non-repeat region

(B)

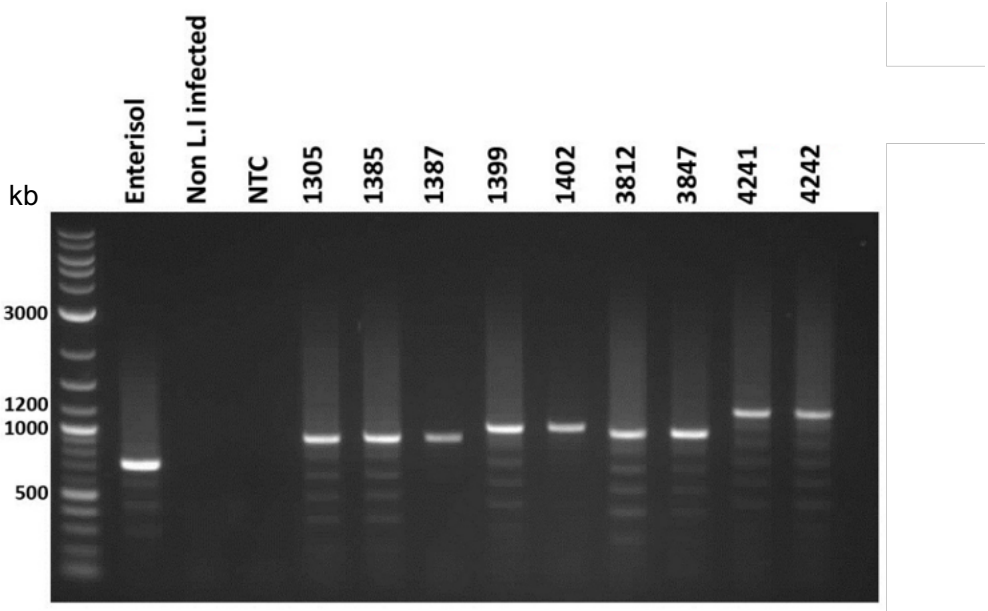

kb
Enterisol
Non L.I infected
NTC
1305
1385
1387
1399
1402
3812
3847
4241
4242
3000
1200
1000
500

Figure 4.2 Variation of *latB* passenger domains among nine clinical *L. intracellularis* isolates. (A) Schematic diagram of PCR primers targeting non-repeat regions flanking *latB* passenger domain. (B) PCR analysis of faecal DNA derived from pigs diagnosed with porcine PE. Amplification was performed using primers targeting non-repeat units flanking the passenger domain of *latB*. Amplified products were resolved on 1 % (w/v) agarose gel. *L. intracellularis* Enterisol® Ileitis vaccine strain was used as positive control, a *L. intracellularis* PCR-negative pig faecal DNA sample and non-template control (NTC) were used as negative controls. Bands from the agarose gel were excised and Sanger sequenced. Sequence analysis were performed using DNASTAR Lasergene® software. Variation in the number of repeating units in the LatB passenger domain among the *L. intracellularis* clinical isolates are displayed in Table 4.2.

Table 4.2 Variation of latB passenger domain in clinical isolates.

Sample	Passenger domain repeat length (aa)	No. long repeat units	No. short repeat units
Enterisol® Ileitis	204	3.5	2
Herd			
1305	269	5.5	1
1385	269	5.5	1
1387	269	5.5	1
1399	298	5.5	2
1402	346	6.5	2
3812	281	5.5	1
3847	281	5.5	1
4241	346	6.5	2
4242	439	8.5	2
LR189	251	4.5	2

Table 4.3 Variation of LatB passenger domain in the cell passaged isolates.

Strain ID	Passaged date	No. passages	Passenger domain repeat length (aa)	No. long repeat units	No. short repeat units
LR189/03	06.03.03	7	298	5.5	2
LR189/03	16.04.03	28	298	5.5	2
LR189/02	03.04.02	12	346	6.5	2
LR189/02	09.04.02	13	298	5.5	2
LR189/02	08.04.02	18	298	5.5	2
LR189/02	02.07.02	23	251	4.5	2
LR189/5/831	11.06.97	6	298	5.5	2
LR189/5/831	25.06.97	8	298	5.5	2
916/91	12.10.94	9	204	3.5	2
916/91	19.10.94	11	204	3.5	2
916/91	14.11.94	14	204	3.5	2
51/89	17.11.93	5	204	3.5	2
51/89	22.11.93	6	204	3.5	2
1482/89	18.05.94	20	439	8.5	2

4.4.3 LatB is expressed during *L. intracellularis* infection

Proteins involved in bacterial pathogenesis will be typically expressed *in vivo* during infection. To investigate if LatB is expressed *in vivo*, immunofluorescence staining was performed on ileum tissue sections from pigs experimentally challenged with *L. intracellularis* at 3, 7, 14, 21, 28, 35 and 42 days post infection (DPI). The staining of LatB was carried out using polyclonal antibodies raised against two overlapping long repeat units of the putative autotransporter passenger domain.

To determine if LatB is expressed *in vitro*, Western blot analysis was performed on the cell lysate of Enterisol® Ileitis with detection of two protein bands at around 30 and 70 kDa (Figure 4.3). The faint band detected at 70 kDa corresponds to the estimated protein molecular weight of LatB at 74 kDa. However, the presence of a thick protein band detected at 30 kDa suggests that the protein is being cleaved. Once translocated to the outer membrane, the passenger domain of ATs may remain covalently bound to the translocator or autocatalytically cleaved and secreted (Dautin et al., 2007). In such case, the molecular weight of LatB expressed in Enterisol (minus the translocator domain) was estimated at 45 kDa, which does not correspond to the lower molecular band detected, implying further processing/cleavage of the protein may have taken place.

To determine if LatB is expressed *in vivo*, ileum tissue sections were co-stained for LatB and *L. intracellularis* using a mouse monoclonal anti-VPM53 antibody that targets a uncharacterised membrane protein, commonly used to detect the presence of *L. intracellularis* during routine diagnostics (McOrist et al., 1987, MacIntyre et al., 2003a). IHC staining was performed twice, by which ileum sections presented in

Figure 4.4 and Figure 4.5 are from different animals at each time points. Staining for *L. intracellularis* using anti-VPM53 detected the presence of the bacterium in only one pig, at 14 DPI (Figure 4.4). This is the peak of infection with highest percentage of crypts infected (Figure 4.6) (MacIntyre et al., 2003a). At 14 DPI, LatB appear to co-localise with *L. intracellularis* which suggests the putative autotransporter remains associated with the bacterium during the peak of infection. The presence of LatB can be detected in the apical cytoplasm of epithelial cells, the location where *L. intracellularis* resides upon entry into host cells following escape from the vacuole, and expression can be detected from 3 to 42 DPI (Figure 4.4 and Figure 4.5). Although, LatB was not detected in one pig at 42 DPI (Figure 4.4), it was apparent in the staining of ileum section from a different pig at 42 DPI (Figure 4.5).

Overall, we have demonstrated that LatB is expressed during the infection and resides in the same intracellular location as the bacterium, which may suggest a role in *L. intracellularis* pathogenesis.

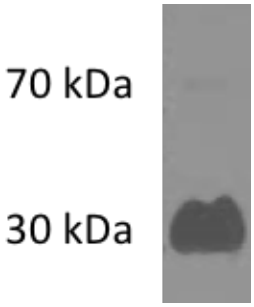


Figure 4.3 Western blot analysis of LatB passenger domain in *L. intracellularis* Enterisol® Ileitis strain cell lysate. Polyclonal antibody generated against two overlapping repeat units of LatB passenger domain was used in Western blot analysis of Enterisol® Ileitis whole cell lysate. The polyclonal antibody detected a thick protein band around 30 kDa and a very faint protein band at 70 kDa. The molecular weight of LatB in Enterisol® Ileitis is estimated to be 74 kDa. Western blot was performed n=3.

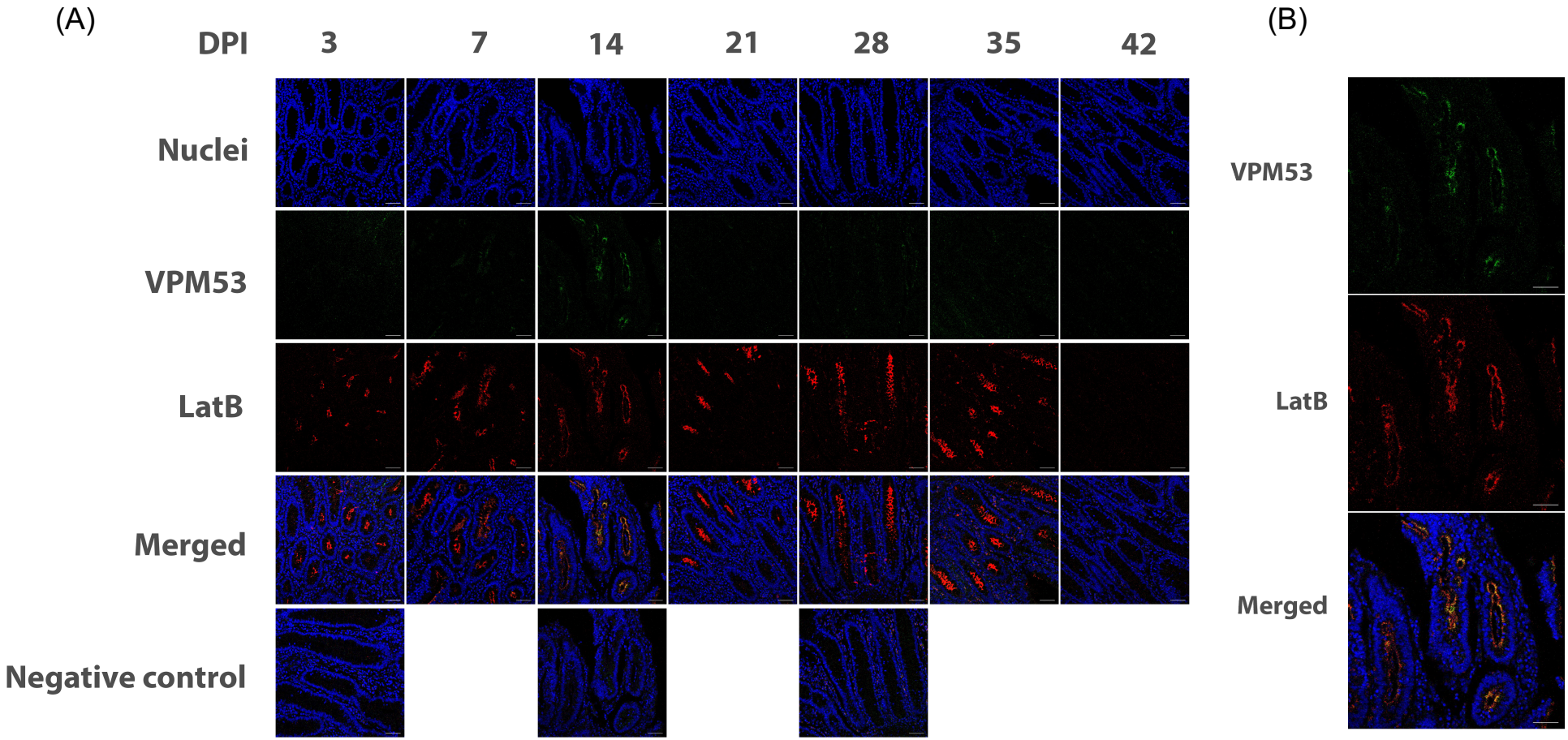

(A)
DPI
3
7
14
21
28
35
42
Nuclei
VPM53
LatB
Merged
Negative control
(B)
VPM53
LatB
Merged

Figure 4.4 IHC of LatB in ileum tissue sections of pigs experimentally challenged with *L. intracellularis* LR189 strain. (A) Ileum tissue sections of pigs experimentally challenged with *L. intracellularis* at 3, 7, 14, 21, 28, 35 and 42 DPI. (B) Ileum sections of 14 DPI enlarged. Nuclei is stained using DAPI (blue), presence of VPM53 is detected using FITC conjugated secondary antibody (green) and the presence of anti-LatB antibody is detected using Alexa Flour®-647-conjugated secondary antibody (red). Merged images represent overlapping of VPM53 and LatB staining, co-localization of the two putative surface proteins at 14 DPI produced yellow/green staining. Negative control images comprised of ileum sections stained without primary antibody. Scale bar represents 50 µm.

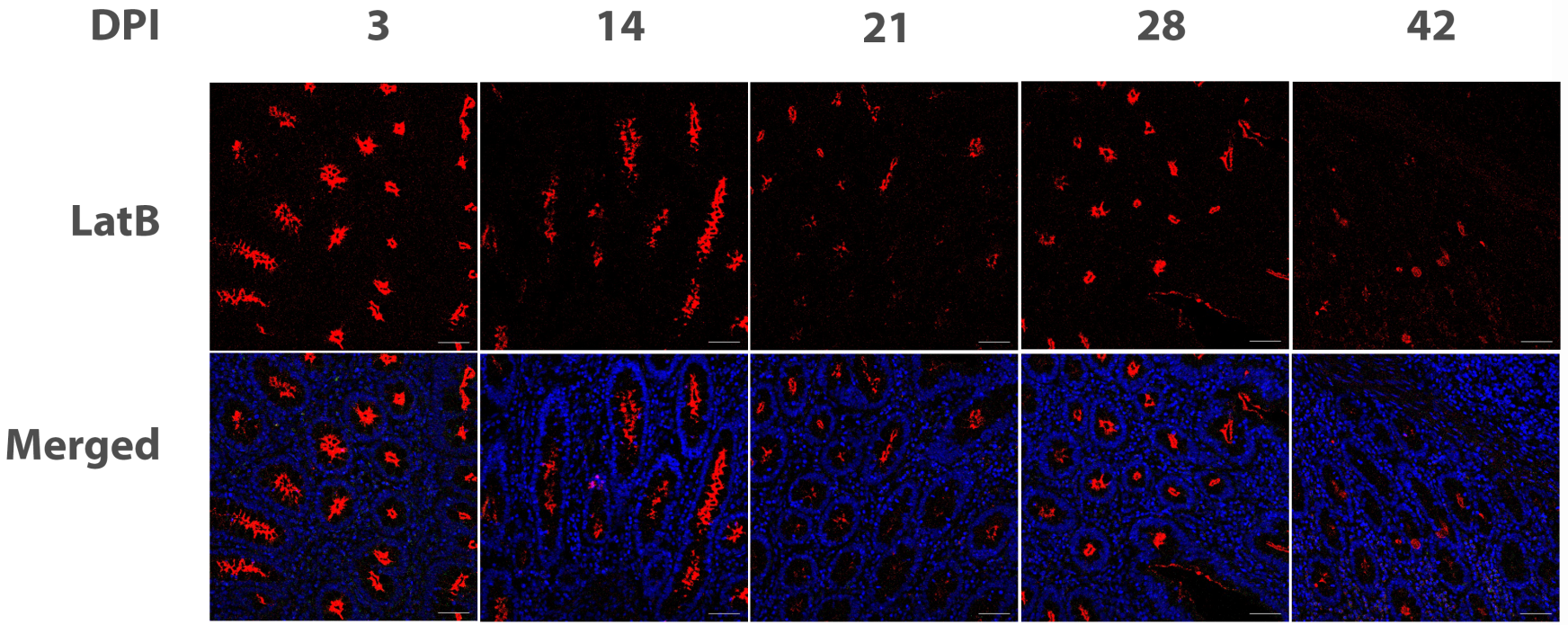

DPI
3
14
21
28
42
LatB
Merged

Figure 4.5 IHC of LatB in ileum tissue sections of pigs experimentally challenged with *L. intracellularis* LR189 strain. Enlarged staining of ileum tissue sections of pigs experimentally challenged with *L. intracellularis* at 3, 14, 21, 28 and 42 DPI. Nuclei is stained by DAPI (blue) and antibody generated to detect a single long repeat unit of the LatB passenger domain is stained in red. The merged image represents overlapping of nuclei and LatB staining. Staining of LatB can be observed in the apical cytoplasm of cells from day 3 of post infection till day 42. Scale bar represents 50 μm.

4.4.4 Generation of recombinant LatB proteins for immunological investigations

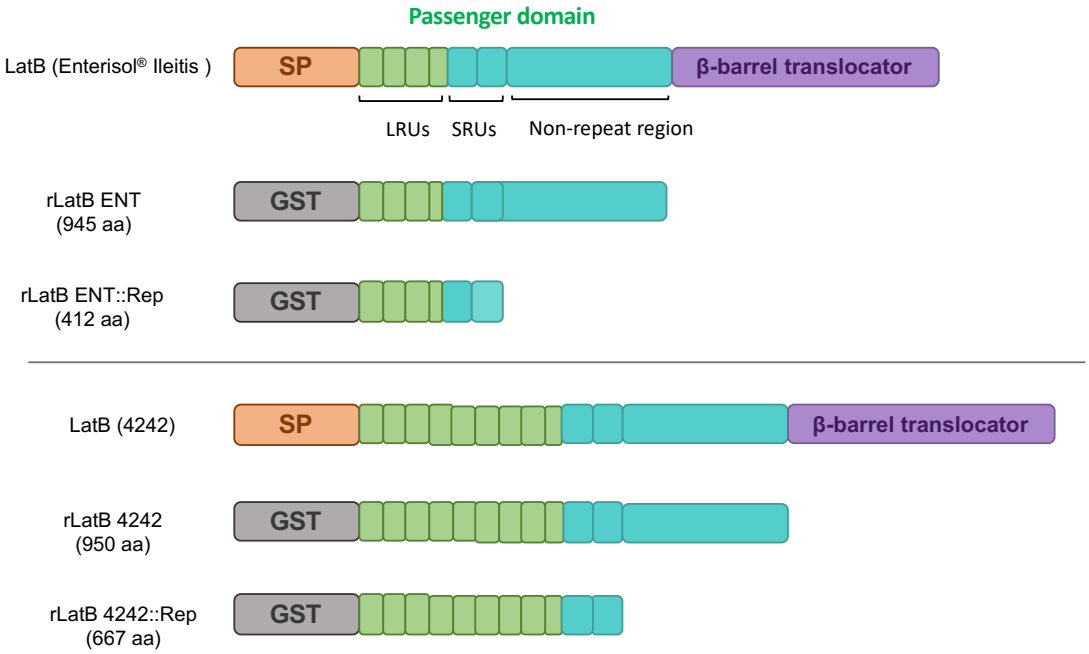
IHC staining of LatB in ileum tissue samples demonstrated that the putative autotransporter is expressed during infection. Previously, transcriptional profiling of *L. intracellularis* infected enterocytes demonstrated a significant up-regulation of genes encoding the major histocompatibility complex class I (MHC-I) protein (Vannucci et al., 2013b), suggesting that *L. intracellularis* derived antigen is presented to the lamina propria, signalling for mucosal immune response (Shao et al., 2005). In order to investigate if LatB is recognised by the host immune system during infection, we expressed recombinant LatB proteins for immunological investigations of LatB and measure its immunoreactivity. To determine if length variation of the passenger domain repeat units produce different serological response, we generated two length variant constructs of LatB. A short variant was cloned from the Enterisol® Ileitis strain, which encodes a passenger domain containing 3.5 long repeat units and 2 short repeat units. A long variant was cloned from the clinical isolate 4242, which encodes a passenger domain containing 8.5 long repeat units and 2 short repeat units.

Recombinant LatB (rLatB) proteins were constructed using the pGEX6P1 vector expressing a GST-tag. The passenger domain of LatB from the Enterisol® Ileitis strain and the clinical 4242 isolate were cloned, expressed in BL21(DE3)/pLysS cells and purified under native conditions, producing rLatB ENT and rLatB 4242 constructs, respectively (Figure 4.6A). The N-terminal signal peptide and conserved C-terminal β-barrel translocator domain was excluded in both constructs, since only the passenger domain is presented on the bacterial surface. Western blot analysis of the purified rLatB constructs showed truncated rLatB ENT products of 65 - 95 kDa and

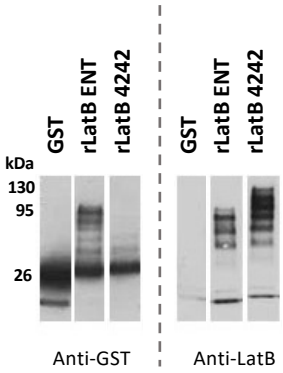
truncated rLatB 4242 products of 65 – 130 kDa, recognised by the anti-LatB antibody (Figure 4.6B). The appearance of distinct protein bands rather than a smear implies specific cleavage rather than non-specific degradation of the recombinant proteins. Constructs of rLatB excluding the 279 aa non-repeat C-terminal region of the passenger domain, rLatB ENT::Rep and rLatB 4242::Rep displayed expression of more stable recombinant proteins of 80 kDa and 100 kDa, respectively (Figure 4.6C). However, expression of rLatB ENT::Rep was detected at almost twice the expected molecular weight of 42 kDa. Recombinant proteins rLatB 4242 and rLatB 4242::Rep cannot be recognised by anti-GST monoclonal antibody, suggesting N-terminal cleavage, or occlusion of the GST tag from both constructs (Figure 4.6B and 4.6C).

In a further attempt to express full length passenger domain of LatB, generation of rLatB proteins were also constructed using the pET28a vector expressing a His_6-tag at the C-terminus. However, expression was weak and Western blot analysis confirmed the presence of truncated products when the autochaperone domain was included in the constructs (data not shown).

(A)


Passenger domain
LatB (Enterisol® Ileitis)
SP
β-barrel translocator
LRUs
SRUs
Non-repeat region
rLatB ENT
(945 aa)
GST
rLatB ENT::Rep
(412 aa)
GST
LatB (4242)
SP
β-barrel translocator
rLatB 4242
(950 aa)
GST
rLatB 4242::Rep
(667 aa)
GST

(B)


GST
rLatB ENT
rLatB 4242
kDa
130
95
26
Anti-GST
GST
rLatB ENT
rLatB 4242
Anti-LatB

(C)

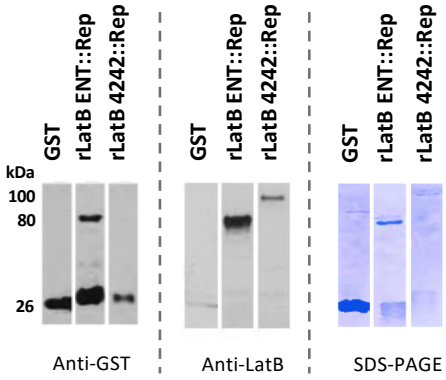

GST
rLatB ENT::Rep
rLatB 4242::Rep
kDa
100
80
26
Anti-GST
GST
rLatB ENT::Rep
rLatB 4242::Rep
Anti-LatB
GST
rLatB ENT::Rep
rLatB 4242::Rep
SDS-PAGE

Figure 4.6 Western blot analysis of recombinant LatB (rLatB) expression. (A) Schematic of native LatB from *L. intracellularis* Enterisol® Ileitis vaccine strain and 4242 clinical isolate, and different rLatB constructs produced in this study. Native LatB consists of a signal peptide (SP), passenger domain containing long repeat units (LRUs), short repeat units (SRUs) and a non-repeat region, followed by a β-barrel translocator domain. LatB from the *L. intracellularis* Enterisol® Ileitis strain contain 3.5 LRUs and 2 SRUs. LatB from the 4242 clinical isolate contain 8.5 LRU and 2 SRUs. All recombinant proteins were GST tagged and purified under native conditions. Predicted molecular weights of each recombinant protein are as follows: rLatB ENT (71 kDa), rLatB ENT::Rep (42 kDa), rLatB 4242 (99 kDa) and rLatB 4242::Rep (71 kDa). (B) Western blot analysis of purified rLatB ENT and rLatB 4242 proteins. (C) Western blot analysis and SDS-PAGE of purified rLatB ENT::Rep and rLatB 4242::Rep proteins. Nitrocellulose membranes were incubated with anti-GST or anti-LatB antibody.

4.4.5 Low levels of immunological response to LatB during *L. intracellularis* infection

To determine if the putative surface protein, LatB, is recognised by the humoral immune response, recombinant proteins corresponding to the passenger domain were used in an ELISA with sera from naturally infected pigs. Two recombinant proteins were tested for antigenicity, rLatB ENT::Rep and rLatB 4242::Rep (Figure 4.6A). Sera from nine pigs were evaluated, three were derived from healthy pigs and six were derived from pigs naturally infected with *L. intracellularis*.

In order to verify the serological status of the samples for *L. intracellularis*, indirect ELISA was performed using wells coated with surface antigens from the Enterisol® Ileitis strain. Detection and quantification of antibodies against *L. intracellularis* from serum samples were measured by A_{450} values. A human serum sample was included in the analysis as a negative control and showed no antibody reactivity with *L. intracellularis* surface antigens (Figure 4.7). Results showed the six serum samples derived from pigs naturally infected with *L. intracellularis* had A_{450} values ranging from 1.57 to 1.77, significantly higher than the three serum samples from uninfected pigs, with p-values < 0.0001 (Figure 4.7). This confirmed that the six positive samples were seropositive for *L. intracellularis*. However, A_{450} values of 0.32 to 0.52 were observed for the three serum samples from uninfected healthy pigs, which could be background-noise due to non-specific binding of antibody.

Before testing the immunorecognition of sera antibodies to the passenger domain of rLatB ENT::Rep and rLatB 4242::Rep recombinant proteins, indirect ELISA was performed with the 26 kDa affinity tag, which showed low A_{450} values of less than 0.14 for all nine serum samples (Figure 4.8A). This ensured a lack of non-specific

interaction for the affinity tag of the recombinant proteins. Indirect ELISA performed with rLatB ENT::Rep recombinant protein revealed that negative sample 1, 2 and 3 have low A_{450} value of 0.30, 0.25 and 0.18, respectively (Figure 4.8B). While the human serum negative control had an A_{450} value of 0.07. When compared to the negative serum sample with the highest A_{450} value, only four out of six positive serum samples had significant levels of antibody (with p-value ≤0.01) recognising this recombinant protein (Figure 4.8B). Sero-recognition for rLatB was variable between individuals, in which positive sample 1 and positive sample 3 had higher A_{450} values of 0.72 and 0.65, respectively. While other positive serum samples had A_{450} values ranged from 0.38 to 0.48 (Figure 4.8B).

Indirect ELISA performed with rLatB 4242::Rep recombinant proteins revealed similar trends in A_{450} values among the serum samples, with the positive sample 1 and positive sample 3 having the highest A_{450} values of 0.52 and 0.35, respectively (Figure 4.8C). However, compared to indirect ELISA perfromed with rLatB ENT::Rep, a lower A_{450} values were observed across all the serum samples, suggesting less antibody recognising the longer variant of rLatB. When compared with negative sample 1, only positive serum sample 1 had statistically higher levels of antibody recognition, with p-value ≤0.0001 (Figure 4.8C). The human serum negative control had an A_{450} value of 0.12.

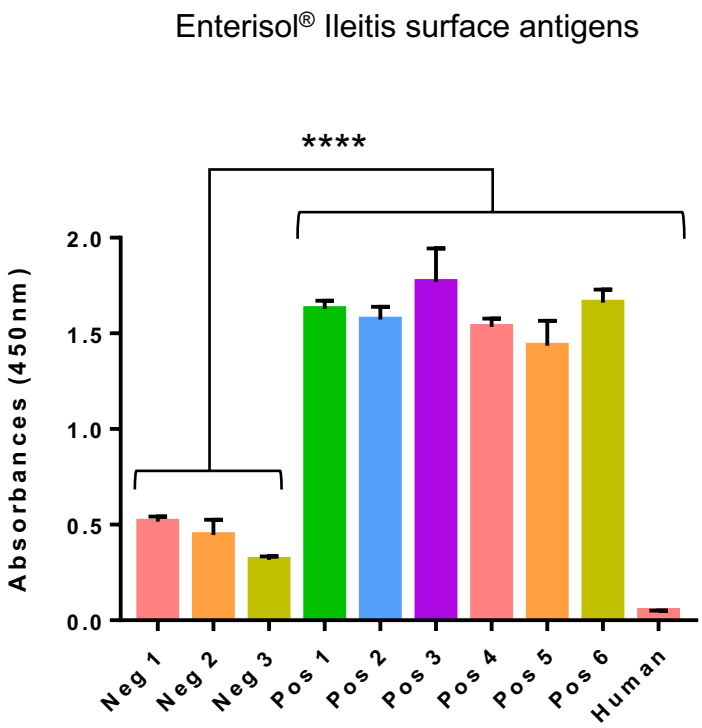
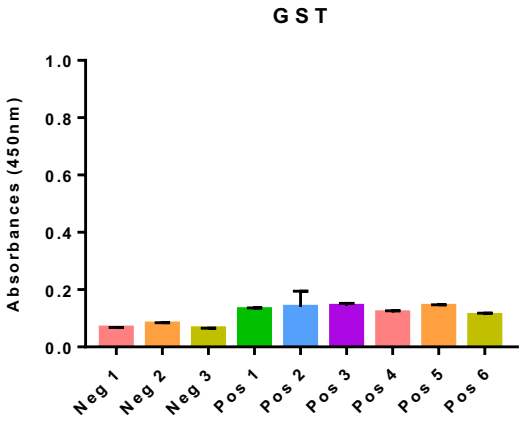
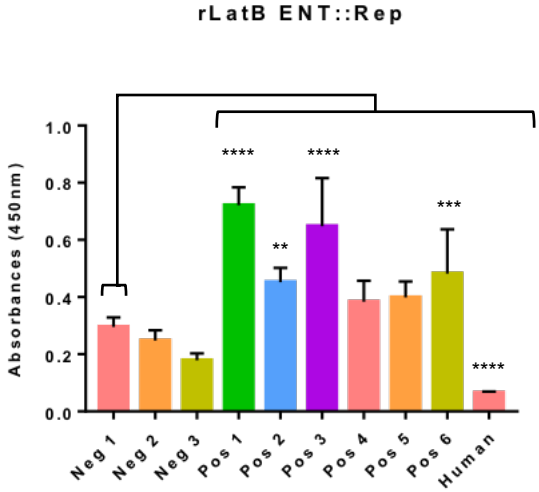


Figure 4.7 Pigs naturally infected with *L. intracellularis* produce antibodies against Enterisol® Ileitis surface antigens. Seroreactivity of nine pig sera and a human serum sample against *L. intracellularis* Enterisol® Ileitis surface antigens were evaluated using indirect ELISA. HRP-conjugated anti-swine IgG antibody was used to detect *L. intracellularis* specific antibody bound to Enterisol® Ileitis surface antigens. Absorbance was measured at 450nm. Results are the mean value of a single experiment performed in triplicate ± standard deviation, n=3. A_{450} value of each positive serum sample and human serum negative control sample were compared to all three negative serum samples individually. Statistical differences were analysed by one-way ANOVA and represented by **** P-value < 0.0001.

(A)


GST
Absorbances (450nm)
1.0
0.8
0.6
0.4
0.2
0.0
Neg 1
Neg 2
Neg 3
Pos 1
Pos 2
Pos 3
Pos 4
Pos 5
Pos 6

(B)


rLatB ENT::Rep
Absorbances (450nm)
1.0
0.8
0.6
0.4
0.2
0.0
****
****
**
***
****
Neg 1
Neg 2
Neg 3
Pos 1
Pos 2
Pos 3
Pos 4
Pos 5
Pos 6
Human

(C)

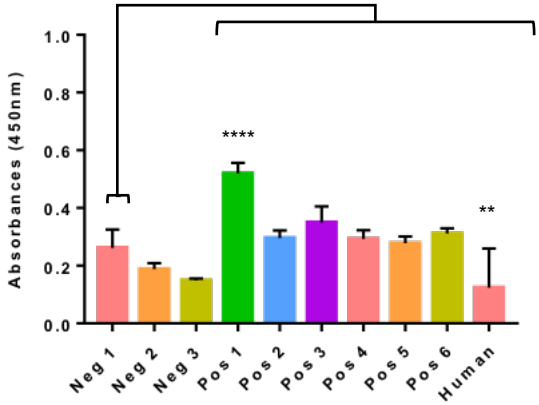

rLatB 4242::Rep
Absorbances (450nm)
1.0
0.8
0.6
0.4
0.2
0.0
****
**
Neg 1
Neg 2
Neg 3
Pos 1
Pos 2
Pos 3
Pos 4
Pos 5
Pos 6
Human

Figure 4.8 Low levels of serum antibodies specific for rLatB in pigs infected with *L. intracellularis*. Indirect ELISA to assess the immune recognition of nine pig sera and a human serum sample for (A) GST-tag, (B) rLatB ENT::Rep and (C) rLatB 4242::Rep recombinant proteins. HRP-conjugated anti-swine IgG antibody was used to detect swine antibody bound to rLatB recombinant constructs. Absorbance was measured at 450nm. Results are the mean value of three independent experiments performed in triplicate ± standard deviation, n=9. A_{450} value of each positive serum sample and human serum negative control sample were compared to negative serum sample 1. Statistical differences were analysed by one-way ANOVA and represented by **** P≤0.0001, *** P≤0.001, ** P≤0.01

4.5 Discussion

Comparative genomic analysis of *L. intracellularis* identified length variation in the *latB* gene, predicted to encode a putative AT protein. ATs comprise a large superfamily of proteins commonly found in pathogenic Gram-negative bacteria and family members have a diverse array of functions, often associated with virulence (Dautin and Bernstein, 2007). We have previously demonstrated that *latB* is part of *L. intracellularis* core genome and all isolates contain a single copy. In the current study, we explored the extent of *latB* variation among *L. intracellularis* isolates and the potential of this putative AT as a therapeutic/diagnostic target for the bacterium.

Sequence analysis of LatB revealed a classical AT motif, comprised of a signal peptide at the N-terminus, followed by a passenger domain and a β-translocator domain at the C-terminus (Figure 4.1A). A BLASTP search of LatB against the NCBI database did not identify conserved domains or functional motifs, with the exception of the conserved AT β-translocator domain. Homologues of LatB were only identified in *L. intracellularis*, all predicted as putative ATs. Although the secondary structure of the LatB β-translocator domain was modelled based on EstA of *P. aeruginosa*, the two protein domains shared low sequence identity of only 14%. Furthermore, a large-scale sequence analysis of ATs has previously showed that the function of the passenger domain does not correspond to β-translocator barrel type, and the two domains occur with a mix and match pattern (Celik et al., 2012). This that suggests even if the β-translocator domain of LatB shares homology with EstA, the passenger domain may not necessarily have the same function.

The passenger domain of ATs corresponds to the functional region of the protein (Henderson et al., 2004). Sequence analysis of the LatB passenger domain revealed the functional region is comprised of long repeating units of 47 aa per unit, short repeating units of 14 aa per unit and a non-repeat region of 274 aa (Figure 4.1A). Sanger sequencing of *latB* passenger domain from *L. intracellularis* isolates demonstrated that variations were limited to the number of long and short repeat units, while sequence of each units remained conserved.

Among clinical isolates, LatB variation was also observed among isolates from the same herd, with the number of long repeating units ranging from 4.5 to 8.5, while the number of short repeating units ranged from 1 to 2 units (Table 4.2). The *L. intracellularis* Enterisol® Ileitis vaccine strain in which attenuation has been achieved through multiple cell culture passages (Kroll et al., 2004). Sequence analysis of LatB passenger domain in Enterisol® Ileitis revealed the shortest LatB passenger domain of 3.5 long repeating units and 2 short repeating units, compared to the clinical isolates (Figure 4.2 and Table 4.2). To determine if the short length variant of *latB* in Enterisol® Ileitis may have been caused by in-frame deletion as a result of passaging, we looked for *latB* variation from *L. intracellularis* isolates cell cultivated between 5 and 28 passages. Our data showed that in most isolates, the length of *latB* was conserved up to passage number 28. (Table 4.3). However, in one strain a reduction of three long repeating units were observed between cell passage number 12 and 23, suggesting that passaging can contribute to loss of repeat units and variation in *latB*. Without genome sequence analysis of the parental strain, we are unable to confirm if the short variant of the *latB* gene in Enterisol® Ileitis is due to strain variation or as a result of deletion due to cell passaging. Nonetheless, the unique short length variant of *latB* may provide a good target as a DIVA (differentiation of infected from vaccinated animals) marker to differentiate from clinical strains. This

could be particularly important, as the major drawback of using an Enterisol® Ileitis vaccine is the inability to differentiate between a clinical field strain and a vaccine strain of *L. intracellualris* in a pig herd. However, testing of more field isolates is essential in order to confirm that the short variant of *latB* is unique to the Enterisol vaccine strain.

To investigate the expression of LatB during infection *in vivo*, we performed IHC staining for LatB on ileum sections of pigs challenged with *L. intracellularis* and euthanised at various time points, originated from a previous experimental infection study by MacIntyre et al (MacIntyre et al., 2003a). Infections were confirmed by the authors through post-mortem examination and IHC staining for *L. intracellularis* at the terminal ileum, using a mouse monoclonal anti-VPM53 antibody (MacIntyre et al., 2003a). IHC confirmed infection in all animals up to 14 DPI, two animals at 21 DPI and only one animal at 28 DPI (MacIntyre et al., 2003a). In the current study, we performed co-staining of LatB with *L. intracellularis* using anti-VPM53 which demonstrated the presence of the bacterium in only one animal at 14 DPI (Figure 4.4). It is worth noting that while in the current study goat anti-mouse IgG FITC was used as secondary reagent for detecting anti-VPM53 antibody, and staining was visualized using a confocal microscopy. In the study by MacIntyre et al, visualization of IHC staining was performed using diaminobenzidine (DAB) procedure (MacIntyre et al., 2003a). Hence, the difference in visualisation method used may have contributed to the disparity in detection of *L. intracellularis* using anti-VPM53 antibody.

IHC staining for LatB demonstrated that the putative AT is expressed during infection *in vivo*, from 3 to 42 DPI (Figure 4.7 and Figure 4.8). During which LatB can be observed within the apical cytoplasm of epithelial cells, the location where

L. intracellularis resides intracellularly after vacuolar escape (McOrist et al., 1995). However, IHC staining for *L. intracellularis* using anti-VPM53 in the terminal ileum section by MacIntyre *et al* demonstrated that infections were cleared in animals 28, 35 and 42 DPI, by which no infected crypts were observed (MacIntyre et al., 2003a). However, qPCR quantification of *L. intracellularis* DNA from the ileal samples demonstrated that infection is still present in few animals at 28, 35 and 42 DPI (Ettinger, 2013). This may suggest that LatB could be marking the presence of *L. intracellularis* within infected cells, in which case may serve as a better diagnostic marker for detection of *L. intracellularis* than VPM53. While co-localisation of LatB with *L. intracellularis* at 14 DPI suggests cell surface expression (Figure 4.4). We are unable to determine whether if LatB remains bound to the β-barrel translocator or not, since the anti-LatB antibody used in the current study targets only two overlapping repeat units of the passenger domain. To confirm localisation of LatB with *L. intracellularis* during infection, more analysis would be needed using antibodies generated against the other regions of the protein.

The expression of LatB during infection suggests that the surface protein could be recognised by the host immune system, through presentation by MHC-I molecule (Shao et al., 2005). To determine if LatB can generate an immunological response, we performed indirect ELISA using rLatB proteins against serum samples from pigs naturally infected with *L. intracellularis*. Ideally, a recombinant protein comprised of the full-length passenger domain, comprising the repeating units and the non-repeat region, would have been analysed. However, efforts to express the full-length passenger domain were unsuccessful. The expression of rLatB cloned from Enterisol® Ileitis and a clinical *L. intracellularis* isolate produced truncated products varying in size between 65 - 95 kDa and 65 – 130 kDa, respectively (Figure 4.6). Conversely, the expression of rLatB proteins excluding the 274 aa non-repeat C-

terminal region of the passenger domain, produced intact recombinant protein constructs (Figure 4.6). This may suggest an autoproteolytic mechanism of the non-repeat region, facilitating self-cleavage. Mechanisms of self-cleavage by protease activity within the passenger domain has been elucidated in several ATs, including IgA1, App and NalP from *Neisseria gonorrhoeae* (Pohlner et al., 1987, Serruto et al., 2003, Turner et al., 2002), and Hap from *Haemophilus influenzae* (Hendrixson et al., 1997). Although the BLASTP search did not identify any conserved protease motifs within the non-repeat passenger domain region of LatB, a putative protease activity may still exist.

Indirect ELISA, performed with serum samples against surface antigen extracted from the *L. intracellularis* Enterisol® Ileitis vaccine strain, showed that pigs positive for *L. intracellularis* demonstrated significantly more immune recognition for surface antigens, compared to uninfected pigs which were negative for the bacterium (Figure 4.7). Low levels of sero-recognition for Enterisol® Ileitis surface antigens were observed, from serum samples derived from uninfected pigs (Figure 4.7). Of note, the serum samples were obtained from field animals and with the high prevalence of *L. intracellularis* across pig-producing countries (Wu et al., 2014, Holyoake et al., 2010, Pascu et al., 2015, Stege et al., 2000), it is likely that the uninfected animals may have had natural exposure to the pathogen. Furthermore, it is feasible that cross-reactivity of sera antibodies for antigens commonly expressed among pig pathogens may contribute to background signals.

Immunological evaluation against recombinant rLatB proteins demonstrated that the putative AT is recognised during infection, with low levels of sero-recognition for rLatB detected from the six *L. intracellularis* seropositive samples (Figure 4.8). We observed a difference in serological response between the long and short rLatB variant, in which

the short variant rLatB ENT::Rep cloned from Enterisol® Ileitis had a higher titre than compared to the long variant rLatB 4242::Rep. The difference in response suggests that pigs may have either been infected with a short LatB variant *L. intracellularis* isolate or have been previously vaccinated with Enterisol® Ileitis, in which case may serve as a potential marker for serological DIVA assay, although this requires further investigation. We also noticed that antibody titre for recognising recombinant rLatB proteins were variable between individuals, in which positive serum sample 1 and positive sample 3 contained a higher titre than compared to the remaining four positive samples from infected pigs (Figure 4.8). This trend was observed for indirect ELISA performed with both the short and long variant of rLatB protein. Variation in antibody titre for *L. intracellularis* among individuals could be influenced by a combination of factors, including the number of *L. intracellularis* organisms infecting an individual, the host immune response and the age of the pig. Furthermore, indirect ELISA with rLatB revealed a low immunological response in the negative serum samples which may suggest antibody produced against the putative AT antigen during exposure to *L. intracellularis* remain, even when there isn't an active infection.

It is important to mention that although measurement of serum IgG may be a useful tool for diagnostic and surveillance purpose, it does not necessarily provide a good measurement for host immune response against infection. Previous studies have showed that increases in IgG titre against *L. intracellularis* does not correlate with protection upon challenge (Cordes et al., 2012). Bearing in mind, that the current analysis is detecting serum IgG antibodies and given the intracellular nature of *L. intracellularis* this type of humoral response will not likely provide protection against infection. Indeed, experimental infection study demonstrated that re-challenged animals responded with significantly higher IFN-γ responses, contributed by $CD8^+$ and $CD4^+$ $CD8^+$ double positive lymphocytes (Cordes et al., 2012). Furthermore,

accumulation of *L. intracellularis* specific IgA at proliferative lesions have been previously described (McOrist et al., 1992, Guedes and Gebhart, 2010). Taken together, cell-mediated immune responses and local IgA are likely to provide protective immunity against *L. intracellularis* infection (Cordes et al., 2012, Guedes and Gebhart, 2003b). Thus, immunological examination of intestinal mucosa IgA for LatB may provide insight to its role in the local intestinal immune response.

In summary, we have shown that *latB* encodes a putative AT protein that exists in multiple length variants among *L. intracellularis* isolates clinical isolates. A short variant of the *latB* gene is identified in the Enterisol® Ileitis vaccine strain, which could be used as a potential DIVA marker to differentiate between vaccine and field clinical isolates. Furthermore, IHC staining for LatB have demonstrated expression of this protein during infection *in vivo*, from 3 to 42 DPI. While the current diagnostic marker VPM53 failed to detect presence of *L. intracellularis* at late stages of infection, detection of LatB at these days may serve as a better diagnostic marker for detection *of L. intracellularis* in IHC. We detected low levels of serum IgG specific for LatB in infected animals, suggesting that the putative surface protein is recognised by the host immune response. Further studies on the functional characterisation of LatB are required in order to investigate the biological role of this protein in *L. intracellularis* pathogenesis.

Chapter 5

General discussion

Proliferative enteropathy caused by *L. intracellularis* is a common intestinal disease of pigs that is endemic with high herd prevalence across major pork producing countries (Wu et al., 2014, Holyoake et al., 2010, Van der Heijden et al., 2004, Lee et al., 2001). This disease impacts animal performance resulting in significant economic loss in the pig farming industry (Jensen, 2006). A recent rise in the number of disease cases have been reported in foals within horse farms across the globe and becoming endemic (Shimizu et al., 2010, Van Den Wollenberg et al., 2011, Pusterla and Gebhart, 2013). Furthermore, PE has been described in a wide range of mammalian species, including wild and domestic animals (Pusterla et al., 2008, Lawson and Gebhart, 2000, Vannucci and Gebhart, 2014, Cooper et al., 1997a). Although the dynamic of the disease has been well characterised in the literature, there is limited understanding of the molecular mechanisms of *L. intracellularis* pathogenesis and the genetic determinants of virulence. This is largely due to its fastidious *in vitro* growth requirements *(Lawson et al., 1993b)*, which make characterisation of the pathogen through conventional laboratory techniques extremely challenging. While the majority of studies have focused on identifying *L. intracellularis* virulence determinants and immunogenetic antigens for vaccine development (Alberdi et al., 2009, McCluskey et al., 2002, Watson et al., 2011, Vannucci et al., 2013d, Won and Lee, 2018), there has been little-to-no studies investigating the genetic diversity of this bacterium. The current work represents the first study to obtain *L. intracellularis* genome through direct sequencing of clinical samples, providing the first report on the population structure and phylogenetic relationship of the pathogen across the globe. This will offer new insights into the evolutionary dynamics of the pathogen, help to inform development of novel therapies and interventions for the control of PE.

To date, only six reference genomes for *L. intracellularis* are deposited in NCBI and have been available since 2006, all of which were obtained from cell cultured

samples. In this work, a total of 20 *L. intracellularis* draft genome sequences were obtained through direct sequencing of faecal and intestinal tissue samples from equine and porcine hosts with PE, with the majority of the genomes acquired from faecal samples. The method developed in the current study therefore provides a practical advantage of being non-invasive, and has great potential to be applied for the future surveillance of *L. intracellularis*. This is essential as rising cases of PE over the years have led to an increase usage of antibiotic and vaccination against *L. intracellularis*. Thus, increased monitoring is required for detecting the emergence of antibiotic resistance or virulent strains.

During industrialisation of agriculture, animal farms moved out of barnyards and into factories, where large groups of animals were selectively bred for high yield traits and housed indoors within a confined space to improve productivity. The shift into intensive livestock production coincided with the emergence (or re-emergence) and spread of infectious diseases (Pulliam et al., 2011, Davies, 2012). It is thought that the expansion of genetically-similar high-density populations may have created changes in selection pressure for higher rate of bacterial transmission promoting proliferation of bacterial population, rapid evolution, and driving pathogen emergence (Weinert et al., 2015, Ezenwa et al., 2006, Engering et al., 2013). Increased wildlife-livestock-human interface due to urbanisation and interspersing of farmland with natural landscapes, may have also contributed to cross-species pathogen transmission and spill-over events (Hassell et al., 2017, Engering et al., 2013). Furthermore, the continuous flow of animals facilitates transmission, rapid spread and maintenance of infectious agents within a population (Hassell et al., 2017). Although in the current work, we were unable to estimate a timeframe for the emergence of *L. intracellularis* porcine lineage (Table 3.5), the rise in porcine PE cases have been previously linked to the industrialisation of swine production in the US (Davies, 2012).

Furthermore, the first case of porcine PE was described in the US in 1931 (Biester and Schwarte, 1931), not long after the industrialization and intensification of the pig production system that occurred in the late 1900s . The lack of geographical clustering observed in this study for some porcine-associated *L. intracellularis* isolates, suggests dissemination of the pathogen across Europe and the Americas (Figure 3.2). We speculate that the spread of *L. intracellularis* is likely to be the result of expanding international livestock trade, as millions of live animals are being transported between countries for purposes of breeding, fattening and slaughter. Since porcine PE cases can be subclinical, high-level pig movements in the livestock trade can easily facilitate spread of the disease without being detected. This finding highlights the need for improved surveillance and control strategies for *L. intracellularis*.

Two clinical manifestations of PE are described in pigs, with the chronic form commonly affecting weaners or young growing animals and the acute form commonly affecting mature grower-finisher pigs (McOrist and Gebhart, 2012). Moreover, subclinical cases have often been recognised in endemic herds by which infected animals may become asymptomatic carriers and intermittently shed *L. intracellularis* for long periods of time (Jacobson et al., 2003). The underlaying cause for variation in clinical presentation and how some animals remain asymptomatic is yet to be determined. In the current work, porcine associated isolates originating from farmed animals presenting various degrees of clinical severity displayed minor genetic variations (Figure 3.2). This observation suggests that different forms of clinical disease may be caused by *L. intracellularis* isolates that are phylogenetically similar. However, due to insufficient clinical data available for the animals, we were unable to associate genetic variants to a clinical phenotype. However, previous experimental infection study demonstrated that animals from the same study group challenged with the same *L. intracellularis* strain at equal dosage, displayed variation in disease

outcome between individuals (MacIntyre et al., 2003a). Thus, we suggest that variation in host immune response upon infection is likely to have an influence on disease outcome. In addition, it is thought that stress, associated with transportation, handling and regrouping of animals, is a predisposing factor to cause a clinical outbreak from a subclinical infection (Bane et al., 2001, Smith et al., 1998, Bae et al., 2013). Stress has an effect on secretion of various hormones, which upon release alter the host immune response and influence susceptibility to infection (Freestone et al., 2008). The alteration of host stress reaction as a signal for growth and pathogenesis has been exploited by several pathogens (Cogan et al., 2006, Vlisidou et al., 2004), including *Salmonella enterica* whereby the release of neuroendocrine during the stress reaction has been shown to promote intracellular proliferation in macrophages (Verbrugghe et al., 2011). Moreover, association between *S. enterica* serovar Typhimurium carriage and *L. intracellularis* infection have been previously identified (Belœil et al., 2004), and vaccination against *L. intracellularis* have shown decreased faecal shedding of *S. enterica* serovar Typhimurium in co-challenged animals (Leite et al., 2018). Suggesting that a weakened state of gut immunity due to co-infection may predispose to colonization by other microbes, contributing to clinical outbreaks of infection. Finally, early experimental infection study revealed that, PE can only be reproduced in gnotobiotic pigs challenged with intestinal homogenate containing *L. intracellularis*, but not when challenged with pure cell culture inoculum (McOrist et al., 1993b). Indicating that disease progression requires interaction with members from the gut microflora.

Cross-species experimental infection studies with rabbit, hamster, mice, porcine and equine hosts have demonstrated species-specific characteristics of *L. intracellularis* infection, by which susceptibility of the animal is dependent on the host-origin of the bacterial isolate (Vannucci et al., 2012b, Sampieri et al., 2013b, Gabardo et al., 2017).

An understanding of the basis for host-specificity will provide important insights into the molecular pathogenesis of *L. intracellularis*, and is central to address questions regarding how the bacterium is able to cross-species barriers and adapt to new hosts. In the present study, phylogenetic analyses of porcine and equine-derived *L. intracellularis* revealed that isolates associated with the two different hosts are phylogenetically distinct and may potentially represent subtypes of *L. intracellularis*. While the phylogenetic tree topology presented in Figure 3.1 demonstrated a host-associated population structure of *L. intracellularis* for porcine and equine hosts, further analysis including more isolates will be required to explore the degree of host association. It would be interesting to examine the phylogenetic relationship of isolates derived from rodents and rabbits, as these are thought to act as biological reservoirs for *L. intracellularis* and responsible for interspecies transmission to porcine and equine hosts (Gabardo et al., 2017, Sampieri et al., 2013a). Allelic profiles generated from MLST and VNTR have previously demonstrated that isolates derived from porcine, equine and rodent hosts all possessed distinct sequence types (Gebhart et al., 2012). A comprehensive analysis of the genetic relationship between these isolates may provide insights into the evolution and host-adaptation of *L. intracellularis* among various hosts, and the role of these animals in the epidemiology of PE.

The equine *L. intracellularis* E40504 strain has previously demonstrated host-specificity for infection in horses and rabbits, but is unable to produce any pathological changes in pigs or hamsters, under experimental infection (Vannucci et al., 2012b, Sampieri et al., 2013b). The present study identified a high level of conservation in gene content among the porcine and equine derived isolates, suggesting that differences in host-specificity among *L. intracellularis* is due to variation in the core genome (Figure 3.4). Indeed, comparative genomic analysis between the E40504 strain and the porcine isolates, revealed high sequence diversification in genes

encoding proteins associated with metabolism, signal transduction, cell wall/ membrane/ envelope biogenesis, replication and repair (Table 3.3). As *L. intracellularis* evolved into porcine and equine adapted lineages, genetic changes may have led to diversification of these genes, resulting in adaptation within the primary host niche that may reduce in fitness for another niche.

Some of the deadliest diseases are caused by bacterial pathogens with extremely low genetic diversity, the most prominent examples being *Bacillus anthracis* (Didelot et al., 2009), *Yersinia pseudotuberculosis* (Ch'ng et al., 2011), *Mycobacterium tuberculosis* complex (Dos Vultos et al., 2008) and *S. enterica* serovar Tyhi (Roumagnac et al., 2006). While some bacteria such as *Staphylococcus aureus,* have evolved into clonal lineages over recent decades (Holden et al., 2013). Others, such as *Yersinia pestis* have evolved over several millennia (Morelli et al., 2010). Clonality is the result of restrictive genetic recombination on an evolutionary scale, by which events that generate genetic diversity occur so rarely that the clonal population structure is not disrupted (Tibayrenc and Ayala, 2017). Such clonality has been observed in porcine associated *L. intracellularis* isolates examined in this study, by which the genome evolves essentially through mutations with no HGT events observed between the isolates (Figure 3.4). However, the underlying evolutionary force resulting in the strict clonality of the porcine associated population remains elusive. It is believed that genetic drift, due to population bottlenecks introduced during host transmission, may have an evolutionary impact in the reduction of diversity among genetically monomorphic obligate intracellular pathogens including *Mycobacterium tuberculosis* and *Chlamydia trachomatis* (Hershberg et al., 2008, Joseph et al., 2012, Kuo et al., 2009). Considering *L. intracellularis* share a similar intracellular lifestyle, the effect of random genetic drift will probably have a profound effect in its evolution. The lack of genetic diversity at a species population level may

also be the result of periodic selection, by which beneficial mutations are selected for resulting in clonal replacement (Smith and Haigh, 1974). However, this theory has been argued against for genetically monomorphic pathogens (Tibayrenc and Ayala, 2017). As observed in *S. enterica* serovar Tyhi, by which selection for resistance for fluoroquinolones resulted in clonal expansion of the H58 haplotype containing mutation in the *gyrA* gene, but did not lead to the elimination of other haplotypes in the population without the *gyrA* mutation (Roumagnac et al., 2006). The observation that clonal populations tend to stay clonal, is indicative that clonality is a stable trait in any given population (Shapiro, 2016). Tibayrenc and Ayala proposed that clonal pathogens may have evolved mechanisms to restrain recombination as an evolutionary strategy to adapt to parasitism, by preventing disruption of favourable multilocus association (Tibayrenc and Ayala, 2017). In the current study, we were constrained by the limited number of *L. intracellularis* genomes which impeded us from inferring the role of natural selection and genetic drift in shaping the molecular evolution of *L. intracellularis*. Thus, future studies should include globally representative isolates that capture an accurate population framework of *L. intracellularis*.

Although the current study has provided new insights to the population structure of *L. intracellularis*, it is important to note that samples included were biased towards isolates derived from clinical disease cases, which may only represent the minority of the population. This may potentially underestimate the diversity of the population and overestimate the extent of *L. intracellularis* clonality. Thus, examination of isolates from subclinical and clinical apparent cases will be required to fully assess the general population of *L. intracellularis*, and elucidate predisposing factors contributing to disease outbreak to inform better control strategies. Furthermore, although the current study has investigated *L. intracellularis* genetic diversity between hosts, we did not

explore its genetic diversity within-host. Capturing such diversity is important for understanding the evolutionary dynamics of the bacterial pathogen during host colonization, a key factor in determining the potential for host-adaptation, response during treatments and identification of mixed infections (Didelot et al., 2016). However, this requires sequencing multiple isolates sampled from the same host, which cannot be achieved through metagenomic sequencing, as current method of *de novo* metagenome assembly is unable to recover single bacterial isolate genome.

To date, live attenuated and inactivated bacterin-based vaccines are available for prophylactic use for the control of *L. intracellularis* (Roerink et al., 2018, McOrist and Smits, 2007). Despite the disadvantage that comes with using an attenuated live vaccine, it is generally preferred over a killed vaccine for protection against intracellular pathogens, due its ability to induce an appropriate T cell-mediated immune response (Griffiths and Khader, 2014). In the current work, comparative analysis between Enterisol® Ileitis and a clinical isolate phylogenetically closest to Enterisol® Ileitis, revealed only limited genetic variation. This finding, plus the lack of understanding of the molecular basis for Enterisol® Ileitis attenuation, raises safety concerns, as few mutations would be required to promote reversion to a pathogenic form. Furthermore, immunity provided by Enterisol® Ileitis is suboptimal and this use of partially effective vaccines in intensive farming system may drive selection of *L. intracellularis* with increase virulence (Gandon et al., 2001). Such vaccine-driven selection is thought to have contributed to the increased virulence of *Bordetella pertussis* and *Plasmodium chabaudi* (van Boven et al., 2005, Octavia et al., 2010, Mackinnon et al., 2008), and the use of Enterisol® Ileitis should be closely monitored.

A recently developed *Salmonella* based vaccine was able to elicit specific humoral and cell-mediated immunity against *L. intracellularis* and *S. enterica* serovar

Typhimurium, conferring protection in a murine challenge model (Park et al., 2018). This vaccine is comprised of live attenuated *S.* Typhimurium expressing and secreting *L. intracellularis* proteins OptA, OptB, LfliC and Lhly, by which OptA and OptB are immunogenic domains of a putative AT protein, LatA (Watson et al., 2011). The ubiquity of AT proteins among Gram-negative pathogens and their role in virulence make them an attractive target for vaccine development (Wells et al., 2007). In the current work, we performed immunological investigation of LatB, a putative *L. intracellularis* AT, which demonstrated immunorecognition in infected animals (Figure 4.8), although the immunogenicity of LatB is yet to be determined. Expression of LatB was observed throughout infection, at the site of *L. intracellularis* infection in porcine ileum (Figure 4.3 and 4.4), suggesting this may serve as a potential vaccine target. Incorporation of the LatB short repeat unit as part of a vectored vaccine or as an adjuvant subunit vaccine for use in conjunction with inactivated vaccines, may serve as a safer and more efficacious alternative to Enterisol® Ileitis.

The current work has demonstrated the potential of metagenomic sequencing to investigate the population genomics of an obligate intracellular pathogen, allowing improved understanding of its evolutionary dynamics, identification of the genetic basis for host association and enabled investigation of a potential diagnostic marker (Figure 5.1). The non-invasive sampling strategy used in this study provides a convenient way to collect samples, and is the starting point for future large-scale genomic studies and disease surveillance of *L. intracellularis*, as well as other fastidious pathogens. Furthermore, microbiome data acquired from metagenomic sequencing may help us to unravel the complex interaction between *L. intracellularis* and members of the gut microflora to cause clinical disease, and understand the key factors leading to PE. Development of novel sequencing technologies such as the

MinION has enabled metagenomic sequencing to become widely accessible in laboratories across the world. In the coming years, as the field of metagenomics matures, such methods can be regularly applied in research centres and public health laboratories to facilitate the investigation of novel microbes responsible for disease whose aetiology is unknown.

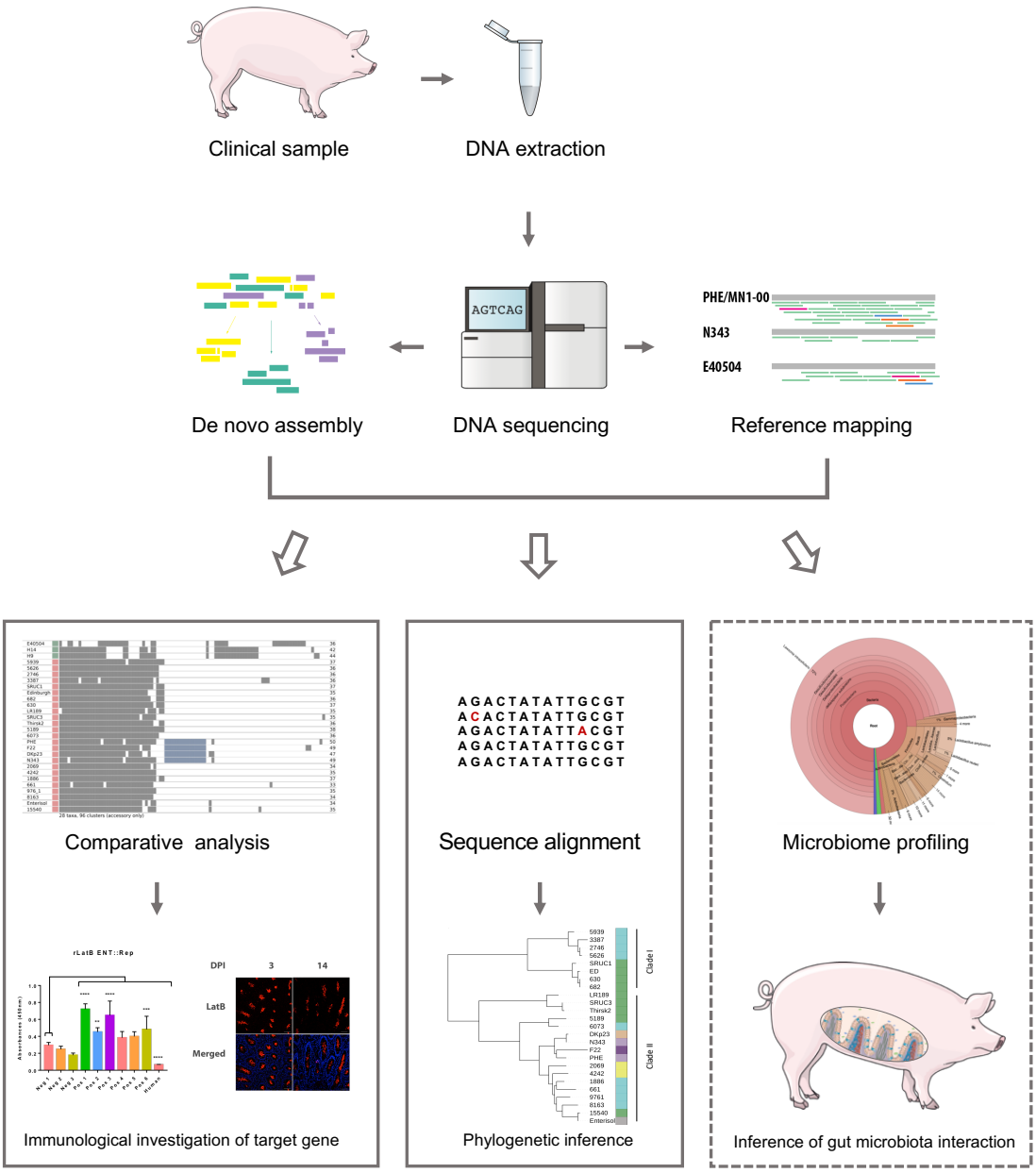


Figure 5.1 Metagenomic sequencing of clinical samples for investigation of population and functional genomics of *L. intracellularis*. In the current work *L. intracellularis* genomes were obtained through direct sequencing of clinical and cell cultured samples. This enabled population genomic analysis for inference of *L. intracellularis* phylogeny, determine genetic basis for host association and identification of potential diagnostic marker. Future studies may utilise microbiome data to examine interaction between *L. intracellularis* with members of gut microbiota to cause clinical disease.

References

ACHTMAN, M. 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual Review of Microbiology,* 62.

ACHTMAN, M. & WAGNER, M. 2008. Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology,* 6**,** 431.

ALBERDI, M. P., WATSON, E., MCALLISTER, G. E., HARRIS, J. D., PAXTON, E. A., THOMSON, J. R. & SMITH, D. G. 2009. Expression by Lawsonia intracellularis of type III secretion system components during infection. *Veterinary Microbiology,* 139**,** 298-303.

ALBERT, M. J., ALAM, K., ANSARUZZAMAN, M., ISLAM, M., RAHMAN, A., HAIDER, K., BHUIYAN, N., NAHAR, S., RYAN, N. & MONTANARO, J. 1992. Pathogenesis of Providencia alcalifaciens-induced diarrhea. *Infection and Immunity,* 60**,** 5017-5024.

ALTSCHUL, S. F., MADDEN, T. L., SCHÄFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25**,** 3389-3402.

AMBROSE, H., GRANEROD, J., CLEWLEY, J., DAVIES, N., KEIR, G., CUNNINGHAM, R., ZUCKERMAN, M., MUTTON, K., WARD, K. & IJAZ, S. 2011. Encephalitis: diagnostic strategy used to establish aetiologies in a prospective cohort of patients in England. *Journal of Clinical Microbiology***,** JCM. 00862-11.

ANANTHARAMAN, K., BROWN, C. T., HUG, L. A., SHARON, I., CASTELLE, C. J., PROBST, A. J., THOMAS, B. C., SINGH, A., WILKINS, M. J. & KARAOZ, U. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications,* 7**,** 13219.

ANDERSSON, S. G. & KURLAND, C. G. 1998. Reductive evolution of resident genomes. *Trends in Microbiology,* 6**,** 263-268.

ANDREWS, S. 2010. FastQC A Quality Control tool for High Throughput Sequence Data

BAE, J., WIELAND, B., SAIT, M., LONGBOTTOM, D., SMITH, D., ALARCON, P. & WHEELHOUSE, N. 2013. Risk factors associated with Lawsonia intracellularis in English pig farms. *The Veterinary Journal,* 197**,** 707-711.

BAKER, K. S., DALLMAN, T. J., FIELD, N., CHILDS, T., MITCHELL, H., DAY, M., WEILL, F.-X., LEFÈVRE, S., TOURDJMAN, M. & HUGHES, G. 2018. Horizontal antimicrobial resistance transfer drives epidemics of multiple Shigella species. *Nature Communications,* 9**,** 1462.

BANE, D. P., NEUMANN, E., GEBHART, C. J., GARDNER, I. A. & NORBY, B. 2001. Porcine proliferative enteropathy: a case-control study in swine herds in the United States. *Journal of Swine Health and Production,* 9**,** 155-160.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S. & PRJIBELSKI, A. D. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology,* 19**,** 455-477.

BARON, E. J. 1997. Bilophila wadsworthia: a unique Gram-negative anaerobic rod. *Anaerobe,* 3**,** 83-86.

BEAULAURIER, J., ZHU, S., DEIKUS, G., MOGNO, I., ZHANG, X.-S., DAVIS-RICHARDSON, A., CANEPA, R., TRIPLETT, E. W., FAITH, J. J. & SEBRA, R. 2018. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nature Biotechnology,* 36**,** 61.

BEER, R. 1973. Studies on the biology of the life-cycle of Trichuris suis Schrank, 1788. *Parasitology,* 67**,** 253-262.

BEITEL, C. W., FROENICKE, L., LANG, J. M., KORF, I. F., MICHELMORE, R. W., EISEN, J. A. & DARLING, A. E. 2014. Strain-and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ,* 2**,** e415.

BELŒIL, P.-A., FRAVALO, P., FABLET, C., JOLLY, J.-P., EVENO, E., HASCOET, Y., CHAUVIN, C., SALVAT, G. & MADEC, F. 2004. Risk factors for Salmonella enterica subsp. enterica shedding by market-age pigs in French farrow-to-finish herds. *Preventive Veterinary Medicine,* 63**,** 103-120.

BENDTSEN, J. D., NIELSEN, H., VON HEIJNE, G. & BRUNAK, S. 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology,* 340**,** 783-795.

BENGTSSON, R. J., MACINTYRE, N., GUTHRIE, J., WILSON, A. D., FINLAYSON, H., MATIKA, O., PONG-WONG, R., SMITH, S. H., ARCHIBALD, A. L. & AIT-ALI, T. 2015. *Lawsonia intracellularis* infection of intestinal crypt cells is associated with specific depletion of secreted MUC2 in goblet cells. *Veterinary Immunology andlimmunopathology,* 168**,** 61-67.

BENJAK, A., AVANZI, C., SINGH, P., LOISEAU, C., GIRMA, S., BUSSO, P., FONTES, A. N. B., MIYAMOTO, Y., NAMISATO, M. & BOBOSHA, K. 2018. Phylogenomics and antimicrobial resistance of the leprosy bacillus *Mycobacterium leprae*. *Nature Communications,* 9**,** 352.

BENTLEY, S. D. & PARKHILL, J. 2015. Genomic perspectives on the evolution and spread of bacterial pathogens. *Proceedings of the Royal Society B,* 282**,** 20150488.

BENZ, I. & SCHMIDT, M. A. 2011. Structures and functions of autotransporter proteins in microbial pathogens. *International Journal of Medical Microbiology,* 301**,** 461-468.

BIERS, E. J., SUN, S. & HOWARD, E. C. 2009. Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Applied and Environmental Microbiology,* 75**,** 2221-2229.

BIESTER, H. & SCHWARTE, L. 1931. Intestinal adenoma in swine. *The American journal of pathology,* 7**,** 175.

BOBAY, L.-M., TRAVERSE, C. C. & OCHMAN, H. 2015. Impermanence of bacterial clones. *Proceedings of the National Academy of Sciences,* 112**,** 8893-8900.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BORDENSTEIN, S. R. & REZNIKOFF, W. S. 2005. Mobile DNA in obligate intracellular bacteria. *Nature Reviews Microbiology,* 3**,** 688.

BORDENSTEIN, S. R. & WERNEGREEN, J. J. 2004. Bacteriophage flux in endosymbionts (Wolbachia): infection frequency, lateral transfer, and recombination rates. *Molecular Biology and Evolution,* 21**,** 1981-1991.

BOUCKAERT, R., HELED, J., KÜHNERT, D., VAUGHAN, T., WU, C.-H., XIE, D., SUCHARD, M. A., RAMBAUT, A. & DRUMMOND, A. J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology,* 10**,** e1003537.

BOUTRUP, T. S., BOESEN, H., BOYE, M., AGERHOLM, J. S. & JENSEN, T. K. 2010. Early pathogenesis in porcine proliferative enteropathy caused by *Lawsonia intracellularis*. *Journal of Comparative Pathology,* 143**,** 101-109.

BOWERS, R. M., KYRPIDES, N. C., STEPANAUSKAS, R., HARMON-SMITH, M., DOUD, D., REDDY, T., SCHULZ, F., JARETT, J., RIVERS, A. R. & ELOE-FADROSH, E. A. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology,* 35**,** 725.

BRADNAM, K. R., FASS, J. N., ALEXANDROV, A., BARANAY, P., BECHNER, M., BIROL, I., BOISVERT, S., CHAPMAN, J. A., CHAPUIS, G. & CHIKHI, R. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience,* 2**,** 10.

BRIGHT, M. & BULGHERESI, S. 2010. A complex journey: transmission of microbial symbionts. *Nature Reviews Microbiology,* 8**,** 218.

BRODRICK, H. J., RAVEN, K. E., HARRISON, E. M., BLANE, B., REUTER, S., TÖRÖK, M. E., PARKHILL, J. & PEACOCK, S. J. 2016. Whole-genome sequencing reveals transmission of vancomycin-resistant *Enterococcus faecium* in a healthcare network. *Genome Medicine,* 8**,** 4.

BROWN, C. T., HUG, L. A., THOMAS, B. C., SHARON, I., CASTELLE, C. J., SINGH, A., WILKINS, M. J., WRIGHTON, K. C., WILLIAMS, K. H. & BANFIELD, J. F. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature,* 523**,** 208.

CAMBRONNE, E. D. & SCHNEEWIND, O. 2005. Bacterial invasins: molecular systems dedicated to the invasion of host tissues. *Concepts in Bacterial Virulence.* Karger Publishers.

CARVER, T., HARRIS, S. R., BERRIMAN, M., PARKHILL, J. & MCQUILLAN, J. A. 2011. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics,* 28**,** 464-469.

CASADEVALL, A. 2008. Evolution of intracellular pathogens. *Annual Review of Microbiology.,* 62**,** 19-33.

CASADEVALL, A. & PIROFSKI, L.-A. 1999. Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. *Infection and Immunity,* 67**,** 3703-3713.

CELIK, N., WEBB, C. T., LEYTON, D. L., HOLT, K. E., HEINZ, E., GORRELL, R., KWOK, T., NADERER, T., STRUGNELL, R. A. & SPEED, T. P. 2012. A bioinformatic strategy for the detection, classification and analysis of bacterial autotransporters. *PLoS One,* 7**,** e43245.

CERASI, M., LIU, J. Z., AMMENDOLA, S., POE, A. J., PETRARCA, P., PESCIAROLI, M., PASQUALI, P., RAFFATELLU, M. & BATTISTONI, A. 2014. The ZupT transporter plays an important role in zinc homeostasis and contributes to *Salmonella enterica* virulence. *Metallomics,* 6**,** 845-853.

CH'NG, S. L., OCTAVIA, S., XIA, Q., DUONG, A., TANAKA, M. M., FUKUSHIMA, H. & LAN, R. 2011. Population structure and evolution of pathogenicity of *Yersinia pseudotuberculosis*. *Applied and Environmental Microbiology,* 77**,** 768-775.

CHAFEE, M. E., FUNK, D. J., HARRISON, R. G. & BORDENSTEIN, S. R. 2009. Lateral phage transfer in obligate intracellular bacteria (Wolbachia): verification from natural populations. *Molecular Biology and Evolution,* 27**,** 501-505.

CHEWAPREECHA, C., HOLDEN, M. T., VEHKALA, M., VÄLIMÄKI, N., YANG, Z., HARRIS, S. R., MATHER, A. E., TUANYOK, A., DE SMET, B. & LE HELLO, S. 2017. Global and regional dissemination and evolution of *Burkholderia pseudomallei*. *Nature Microbiology,* 2**,** 16263.

COGAN, T. A., THOMAS, A. O., REES, L. E., TAYLOR, A. H., JEPSON, M. A., WILLIAMS, P. H., KETLEY, J. & HUMPHREY, T. J. 2006. Norepinephrine increases the pathogenic potential of *Campylobacter jejuni*. *Gut*.

COHAN, F. M. & PERRY, E. B. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Current Biology,* 17**,** R373-R386.

COLLINS, A., FELL, S., PEARSON, H. & TORIBIO, J.-A. 2011. Colonisation and shedding of *Lawsonia intracellularis* in experimentally inoculated rodents and in wild rodents on pig farms. *Veterinary Microbiology,* 150**,** 384-388.

COLLINS, A. & LOVE, R. 2007. Re-challenge of pigs following recovery from proliferative enteropathy. *Veterinary Microbiology,* 120**,** 381-386.

COLLINS, A., LOVE, R., POZO, J., SMITH, S. & MCORIST, S. 2000. Studies on the ex vivo survival of *Lawsonia intracellularis*. *Swine Health and Production,* 8**,** 211-216.

COOPER, D., SWANSON, D. & GEBHART, C. 1997a. Diagnosis of proliferative enteritis in frozen and formalin-fixed, paraffin-embedded tissues from a hamster, horse, deer and ostrich using a *Lawsonia intracellularis*-specific multiplex PCR assay. *Veterinary Microbiology,* 54**,** 47-62.

COOPER, D. M., SWANSON, D. L., BARNS, S. M. & GEBHART, C. J. 1997b. Comparison of the 16S ribosomal DNA sequences from the intracellular agents of proliferative enteritis in a hamster, deer, and ostrich with the sequence of a porcine isolate of *Lawsonia intracellularis*. *International Journal of Systematic and Evolutionary Microbiology,* 47.

CORDES, H., RIBER, U., JENSEN, T. K. & JUNGERSEN, G. 2012. Cell-mediated and humoral immune responses in pigs following primary and challenge-exposure to *Lawsonia intracellularis*. *Veterinary Research,* 43**,** 9.

CROUCHER, N. J., PAGE, A. J., CONNOR, T. R., DELANEY, A. J., KEANE, J. A., BENTLEY, S. D., PARKHILL, J. & HARRIS, S. R. 2014. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research,* 43**,** e15-e15.

CUI, L., NEOH, H.-M., IWAMOTO, A. & HIRAMATSU, K. 2012. Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. *Proceedings of the National Academy of Sciences,* 109**,** E1647-E1656.

DARLING, A. E., MAU, B. & PERNA, N. T. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one,* 5**,** e11147.

DAUTIN, N., BARNARD, T. J., ANDERSON, D. E. & BERNSTEIN, H. D. 2007. Cleavage of a bacterial autotransporter by an evolutionarily convergent autocatalytic mechanism. *The EMBO journal,* 26**,** 1942-1952.

DAUTIN, N. & BERNSTEIN, H. D. 2007. Protein secretion in gram-negative bacteria via the autotransporter pathway. *Annual Review of Microbiology,* 61**,** 89-112.

DAVIES, P. 2012. One world, one health: the threat of emerging swine diseases. A North American perspective. *Transboundary and Emerging Diseases,* 59**,** 18-26.

DESAI, A., MARWAH, V. S., YADAV, A., JHA, V., DHAYGUDE, K., BANGAR, U., KULKARNI, V. & JERE, A. 2013. Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PloS one,* 8**,** e60204.

DEWEY, C. E., COX, B. D., STRAW, B. E., BUSH, E. J. & HURD, S. 1999. Use of antimicrobials in swine feeds in the United States. *Journal of Swine Health and Production,* 7**,** 19-25.

DICK, G. J., ANDERSSON, A. F., BAKER, B. J., SIMMONS, S. L., THOMAS, B. C., YELTON, A. P. & BANFIELD, J. F. 2009. Community-wide analysis of microbial genome sequence signatures. *Genome Biology,* 10**,** 1.

DIDELOT, X., BARKER, M., FALUSH, D. & PRIEST, F. G. 2009. Evolution of pathogenicity in the *Bacillus cereus* group. *Systematic and Applied Microbiology,* 32**,** 81-90.

DIDELOT, X., MÉRIC, G., FALUSH, D. & DARLING, A. E. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics,* 13**,** 256.

DIDELOT, X., WALKER, A. S., PETO, T. E., CROOK, D. W. & WILSON, D. J. 2016. Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology,* 14**,** 150.

DOAN, T., WILSON, M. R., CRAWFORD, E. D., CHOW, E. D., KHAN, L. M., KNOPP, K. A., O'DONOVAN, B. D., XIA, D., HACKER, J. K. & STEWART, J. M. 2016. Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens. *Genome Medicine,* 8**,** 90.

DOS VULTOS, T., MESTRE, O., RAUZIER, J., GOLEC, M., RASTOGI, N., RASOLOFO, V., TONJUM, T., SOLA, C., MATIC, I. & GICQUEL, B. 2008. Evolution and diversity of clonal bacteria: the paradigm of *Mycobacterium tuberculosis*. *PloS one,* 3**,** e1538.

DOUGHTY, E. L., SERGEANT, M. J., ADETIFA, I., ANTONIO, M. & PALLEN, M. J. 2014. Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ,* 2**,** e585.

EIDE, D. J. 2005. The ZIP family of zinc transporters. *Zinc Finger Proteins.* Springer.

ENGERING, A., HOGERWERF, L. & SLINGENBERGH, J. 2013. Pathogen–host–environment interplay and disease emergence. *Emerging Microbes & Infections,* 2**,** e5.

EZENWA, V. O., GODSEY, M. S., KING, R. J. & GUPTILL, S. C. 2006. Avian diversity and West Nile virus: testing associations between biodiversity and infectious

disease risk. *Proceedings of the Royal Society of London B: Biological Sciences,* 273**,** 109-117.

FEEHERY, G. R., YIGIT, E., OYOLA, S. O., LANGHORST, B. W., SCHMIDT, V. T., STEWART, F. J., DIMALANTA, E. T., AMARAL-ZETTLER, L. A., DAVIS, T. & QUAIL, M. A. 2013. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PloS one,* 8**,** e76096.

FINKBEINER, S. R., ALLRED, A. F., TARR, P. I., KLEIN, E. J., KIRKWOOD, C. D. & WANG, D. 2008. Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathogens,* 4**,** e1000011.

FOX, J. G., DEWHIRST, F., FRASER, G., PASTER, B., SHAMES, B. & MURPHY, J. 1994. Intracellular Campylobacter-like organism from ferrets and hamsters with proliferative bowel disease is a *Desulfovibrio sp*. *Journal of Clinical Microbiology,* 32**,** 1229-1237.

FRAZER, M. 2008. Lawsonia intracellularis infection in horses: 2005–2007. *Journal of Veterinary Internal Medicine,* 22**,** 1243-1248.

FREESTONE, P. P., SANDRINI, S. M., HAIGH, R. D. & LYTE, M. 2008. Microbial endocrinology: how stress influences susceptibility to infection. *Trends in Microbiology,* 16**,** 55-64.

FRIEDMAN, M., BEDNÁŘ, V., KLIMEŠ, J., SMOLA, J., MRLÍK, V. & LITERÁK, I. 2008. *Lawsonia intracellularis* in rodents from pig farms with the occurrence of porcine proliferative enteropathy. *Letters in Applied Microbiology,* 47**,** 117-121.

FUNK, D. J., WERNEGREEN, J. J. & MORAN, N. A. 2001. Intraspecific variation in symbiont genomes: bottlenecks and the aphid-Buchnera association. *Genetics,* 157**,** 477-489.

GABARDO, M. D. P., SATO, J. P. H., DANIEL, A. G. D. S., ANDRADE, M. R., PEREIRA, C. E. R., REZENDE, T. P., OTONI, L. V. A., REZENDE, L. A. & GUEDES, R. M. 2017. Evaluation of the involvement of mice (*Mus musculus*) in the epidemiology of porcine proliferative enteropathy. *Veterinary Microbiology,* 205**,** 75-79.

GAGNEUX, S. 2018. Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews Microbiology,* 16**,** 202.

GANDON, S., MACKINNON, M. J., NEE, S. & READ, A. F. 2001. Imperfect vaccines and the evolution of pathogen virulence. *Nature,* 414**,** 751.

GAUDRIAULT, S., PAGES, S., LANOIS, A., LAROUI, C., TEYSSIER, C., JUMAS-BILAK, E. & GIVAUDAN, A. 2008. Plastic architecture of bacterial genome revealed by comparative genomics of Photorhabdus variants. *Genome Biology,* 9**,** R117.

GEBHART, C., KELLEY, M. & CHANDER, Y. 2012. Molecular typing of equine *Lawsonia intracellularis* isolates. *Journal of Equine Veterinary Science,* 32**,** S32-S33.

GEBHART, C. J., BARNS, S. M., MCORIST, S., LIN, G.-F. & LAWSON, G. H. 1993. Ileal symbiont intracellularis, an obligate intracellular bacterium of porcine intestines showing a relationship to *Desulfovibrio* species. *International Journal of Systematic and Evolutionary Microbiology,* 43**,** 533-538.

GILL, S. R., POP, M., DEBOY, R. T., ECKBURG, P. B., TURNBAUGH, P. J., SAMUEL, B. S., GORDON, J. I., RELMAN, D. A., FRASER-LIGGETT, C. M. & NELSON, K. E. 2006.

Metagenomic analysis of the human distal gut microbiome. *Science,* 312**,** 1355-1359.

GOERING, R. V. 2010. Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. *Infection, Genetics and Evolution,* 10**,** 866-875.

GOGOLEWSKI, R., COOK, R. & BATTERHAM, E. 1991. Suboptimal growth associated with porcine intestinal adenomatosis in pigs in nutritional studies. *Australian Veterinary Journal,* 68**,** 406-408.

GRASS, G., WONG, M. D., ROSEN, B. P., SMITH, R. L. & RENSING, C. 2002. ZupT is a Zn (II) uptake system in *Escherichia coli*. *Journal of Bacteriology,* 184**,** 864-866.

GRAY, J. P. & HERWIG, R. P. 1996. Phylogenetic analysis of the bacterial communities in marine sediments. *Applied and Environmental Microbiology,* 62**,** 4049-4059.

GREENBLUM, S., TURNBAUGH, P. J. & BORENSTEIN, E. 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences,* 109**,** 594-599.

GRIFFITHS, K. L. & KHADER, S. A. 2014. Novel vaccine approaches for protection against intracellular pathogens. *Current Opinion in Immunology,* 28**,** 58-63.

GUEDES, R. 2004. Update on epidemiology and diagnosis of porcine proliferative enteropathy. *Journal of Swine Health and Production,* 12**,** 134-138.

GUEDES, R. & GEBHART, C. J. 2010. Evidence of cell-mediated immune response and specific local mucosal immunoglobulin (Ig) A production against *Lawsonia intracellularis* in experimentally infected swine. *Canadian Journal of Veterinary Research,* 74**,** 97-101.

GUEDES, R. M. & GEBHART, C. J. 2003a. Comparison of intestinal mucosa homogenate and pure culture of the homologous *Lawsonia intracellularis* isolate in reproducing proliferative enteropathy in swine. *Veterinary Microbiology,* 93**,** 159-166.

GUEDES, R. M. & GEBHART, C. J. 2003b. Onset and duration of fecal shedding, cell-mediated and humoral immune responses in pigs after challenge with a pathogenic isolate or attenuated vaccine strain of *Lawsonia intracellularis*. *Veterinary Microbiology,* 91.

GUREVICH, A., SAVELIEV, V., VYAHHI, N. & TESLER, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-1075.

GYLES, C. L., PRESCOTT, J. F., SONGER, J. G. & THOEN, C. O. 2008. *Pathogenesis of Bacterial Infections in Animals*, John Wiley & Sons.

HARRIS, S. R., CARTWRIGHT, E. J., TÖRÖK, M. E., HOLDEN, M. T., BROWN, N. M., OGILVY-STUART, A. L., ELLINGTON, M. J., QUAIL, M. A., BENTLEY, S. D. & PARKHILL, J. 2013. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *The Lancet Infectious Diseases,* 13**,** 130-136.

HARRIS, S. R., CLARKE, I. N., SETH-SMITH, H. M., SOLOMON, A. W., CUTCLIFFE, L. T., MARSH, P., SKILTON, R. J., HOLLAND, M. J., MABEY, D. & PEELING, R. W. 2012. Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies

phylogenetic relationships masked by current clinical typing. *Nature Genetics,* 44**,** 413.

HASMAN, H., SAPUTRA, D., SICHERITZ-PONTEN, T., LUND, O., SVENDSEN, C. A., FRIMODT-MØLLER, N. & AARESTRUP, F. M. 2013. Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples. *Journal of Clinical Microbiology***,** JCM. 02452-13.

HASSELL, J. M., BEGON, M., WARD, M. J. & FÈVRE, E. M. 2017. Urbanization and disease emergence: Dynamics at the wildlife–livestock–human interface. *Trends in Ecology & Evolution,* 32**,** 55-67.

HENDERSON, I. R., NAVARRO-GARCIA, F., DESVAUX, M., FERNANDEZ, R. C. & ALA'ALDEEN, D. 2004. Type V protein secretion pathway: the autotransporter story. *Microbiology and Molecular Biology Reviews,* 68**,** 692-744.

HENDRIXSON, D. R., DE LA MORENA, M. L., STATHOPOULOS, C. & ST GEME III, J. W. 1997. Structural determinants of processing and secretion of the *Haemophilus influenzae* Hap protein. *Molecular Microbiology,* 26**,** 505-518.

HERSHBERG, R., LIPATOV, M., SMALL, P. M., SHEFFER, H., NIEMANN, S., HOMOLKA, S., ROACH, J. C., KREMER, K., PETROV, D. A. & FELDMAN, M. W. 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biology,* 6**,** e311.

HOLDEN, M. T., HSU, L.-Y., KURT, K., WEINERT, L. A., MATHER, A. E., HARRIS, S. R., STROMMENGER, B., LAYER, F., WITTE, W. & DE LENCASTRE, H. 2013. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Research*.

HOLYOAKE, P., EMERY, D., GONSALVES, J., DONAHOO, M. & COLLINS, A. 2010. Prevalence of antibodies to *Lawsonia intracellularis* in pig herds in Australia. *Australian Veterinary Journal,* 88**,** 186-188.

HONGO, J. A., CASTRO, G. M., CINTRA, L. C., ZERLOTINI, A. & LOBO, F. P. 2015. POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC Genomics,* 16**,** 567.

HUAN, Y. W., BENGTSSON, R. J., MACINTYRE, N., GUTHRIE, J., FINLAYSON, H., SMITH, S. H., ARCHIBALD, A. L. & AIT-ALI, T. 2017. *Lawsonia intracellularis* exploits β-catenin/Wnt and Notch signalling pathways during infection of intestinal crypt to alter cell homeostasis and promote cell proliferation. *PloS one,* 12**,** e0173782.

HUG, L. A., BAKER, B. J., ANANTHARAMAN, K., BROWN, C. T., PROBST, A. J., CASTELLE, C. J., BUTTERFIELD, C. N., HERNSDORF, A. W., AMANO, Y. & ISE, K. 2016. A new view of the tree of life. *Nature Microbiology,* 1**,** 16048.

HUSMEIER, D. 2005. Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics,* 21**,** ii166-ii172.

HWANG, J.-M., SEO, M.-J. & YEH, J.-Y. 2017. *Lawsonia intracellularis* in the feces of wild rodents and stray cats captured around equine farms. *BMC Veterinary Research,* 13**,** 233.

IQBAL, Z., CACCAMO, M., TURNER, I., FLICEK, P. & MCVEAN, G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics,* 44**,** 226.

JACOBSON, M., AF SEGERSTAD, C. H., GUNNARSSON, A., FELLSTRÖM, C., DE VERDIER KLINGENBERG, K., WALLGREN, P. & JENSEN-WAERN, M. 2003. Diarrhoea in the growing pig–a comparison of clinical, morphological and microbial findings between animals from good and poor performance herds. *Research in Veterinary Science,* 74**,** 163-169.

JACOBSON, M., ASPAN, A., KÖNIGSSON, M. H., AF SEGERSTAD, C. H., WALLGREN, P., FELLSTRÖM, C., JENSEN-WAERN, M. & GUNNARSON, A. 2004. Routine diagnostics of *Lawsonia intracellularis* performed by PCR, serological and post mortem examination, with special emphasis on sample preparation methods for PCR. *Veterinary Microbiology,* 102**,** 189-201.

JACOBSON, M., FELLSTRÖM, C. & JENSEN-WAERN, M. 2010. Porcine proliferative enteropathy: an important disease with questions remaining to be solved. *The Veterinary Journal,* 184**,** 264-268.

JACOBSON, M., GERTH LÖFSTEDT, M., HOLMGREN, N., LUNDEHEIM, N. & FELLSTRÖM, C. 2005. The prevalences of *Brachyspira* spp. and *Lawsonia intracellularis* in Swedish piglet producing herds and wild boar population. *Journal of Veterinary Medicine, Series B,* 52**,** 386-391.

JASNI, S., MCORIST, S. & LAWSON, G. 1994. Reproduction of proliferative enteritis in hamsters with a pure culture of porcine ileal symbiont intracellularis. *Veterinary Microbiology,* 41**,** 1-9.

JENSEN, H. M. L. 2006. Health management with reduced antibiotic use—experiences of a Danish pig vet. *Animal Biotechnology,* 17**,** 189-194.

JOSEPH, S. J., DIDELOT, X., ROTHSCHILD, J., DE VRIES, H. J., MORRÉ, S. A., READ, T. D. & DEAN, D. 2012. Population genomics of *Chlamydia trachomatis*: insights on drift, selection, recombination, and population structure. *Molecular Biology and Evolution,* 29**,** 3933-3946.

KAJAVA, A. V. & STEVEN, A. C. 2006. The turn of the screw: variations of the abundant β-solenoid motif in passenger domains of type V secretory proteins. *Journal of Structural Biology,* 155**,** 306-315.

KALYAANAMOORTHY, S., MINH, B. Q., WONG, T. K., VON HAESELER, A. & JERMIIN, L. S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods,* 14**,** 587.

KANG, D. D., FROULA, J., EGAN, R. & WANG, Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ,* 3**,** e1165.

KELLEY, L. A., MEZULIS, S., YATES, C. M., WASS, M. N. & STERNBERG, M. J. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols,* 10**,** 845.

KROLL, J. J., ROOF, M. B., HOFFMAN, L. J., DICKSON, J. S. & HARRIS, D. H. 2005. Proliferative enteropathy: a global enteric disease of pigs caused by *Lawsonia intracellularis*. *Animal Health Research Reviews,* 6**,** 173-197.

KROLL, J. J., ROOF, M. B. & MCORIST, S. 2004. Evaluation of protective immunity in pigs following oral administration of an avirulent live vaccine of *Lawsonia intracellularis*. *American Journal of Veterinary Research,* 65**,** 559-565.

KRZYWINSKI, M. I., SCHEIN, J. E., BIROL, I., CONNORS, J., GASCOYNE, R., HORSMAN, D., JONES, S. J. & MARRA, M. A. 2009. Circos: an information aesthetic for comparative genomics. *Genome Research*.

KUJIRAOKA, M., KURODA, M., ASAI, K., SEKIZUKA, T., KATO, K., WATANABE, M., MATSUKIYO, H., SAITO, T., ISHII, T. & KATADA, N. 2017. Comprehensive diagnosis of bacterial infection associated with acute cholecystitis using metagenomic approach. *Frontiers in Microbiology,* 8**,** 685.

KUO, C.-H., MORAN, N. A. & OCHMAN, H. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Research,* 19**,** 1450-1454.

LA, T., COLLINS, A., PHILLIPS, N., OKSA, A. & HAMPSON, D. 2006. Development of a multiplex-PCR for rapid detection of the enteric pathogens *Lawsonia intracellularis*, *Brachyspira hyodysenteriae*, and *Brachyspira pilosicoli* in porcine faeces. *Letters in Applied Microbiology,* 42**,** 284-288.

LAVOIE, J., DROLET, R., PARSONS, D., LEGUILLETTE, R., SAUVAGEAU, R., SHAPIRO, J., HOULE, L., HALLE, G. & GEBHART, C. 2000. Equine proliferative enteropathy: a cause of weight loss, colic, diarrhoea and hypoproteinaemia in foals on three breeding farms in Canada. *Equine Veterinary Journal,* 32**,** 418-425.

LAWSON, G. & GEBHART, C. 2000. Proliferative enteropathy. *Journal of Comparative Pathology,* 122**,** 77-100.

LAWSON, G., MACKIE, R., SMITH, D. & MCORIST, S. 1995. Infection of cultured rat enterocytes by Ileal symbiont intracellularis depends on host cell function and actin polymerisation. *Veterinary Microbiology,* 45**,** 339-350.

LAWSON, G., MCORIST, S., JASNI, S. & MACKIE, R. 1993a. Intracellular bacteria of porcine proliferative enteropathy: cultivation and maintenance in vitro. *Journal of Clinical Microbiology,* 31**,** 1136-1142.

LAWSON, G. & ROWLAND, A. 1974. Intestinal adenomatosis in the pig: a bacteriological study. *Research in Veterinary Science,* 17**,** 331.

LAWSON, G., ROWLAND, A. & MACINTYRE, N. 1985. Demonstration of a new intracellular antigen in porcine intestinal adenomatosis and hamster proliferative ileitis. *Veterinary Microbiology,* 10**,** 303-313.

LAWSON, G. H., MCORIST, S., JASNI, S. & MACKIE, R. A. 1993b. Intracellular bacteria of porcine proliferative enteropathy: cultivation and maintenance in vitro. *Journal of Clinical Microbiology,* 31.

LEE, S. W., KIM, T. J., PARK, S. Y., SONG, C. S., CHANG, H. K., YEH, J. K., PARK, H. I. & LEE, J. B. 2001. Prevalence of porcine proliferative enteropathy and its control with tylosin in Korea. *Journal of Veterinary Science,* 2**,** 209-212.

LEITE, F. L., SINGER, R. S., WARD, T., GEBHART, C. J. & ISAACSON, R. E. 2018. Vaccination Against *Lawsonia intracellularis* Decreases Shedding of Salmonella enterica serovar Typhimurium in Co-Infected Pigs and Alters the Gut Microbiome. *Scientific Reports,* 8**,** 2857.

LI, D., LIU, C.-M., LUO, R., SADAKANE, K. & LAM, T.-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* 31**,** 1674-1676.

LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v1*.

LI, H., GAO, H., MENG, H., WANG, Q., LI, S., CHEN, H., LI, Y. & WANG, H. 2018. Detection of pulmonary infectious pathogens from lung biopsy tissues by metagenomic next-generation sequencing. *Frontiers in Cellular and Infection Microbiology,* 8**,** 205.

LI, L., STOECKERT, C. J. & ROOS, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research,* 13**,** 2178-2189.

LINDSEY, D., MULLIN, D. & WALKER, J. 1989. Characterization of the cryptic lambdoid prophage DLP12 of *Escherichia coli* and overlap of the DLP12 integrase gene with the tRNA gene argU. *Journal of Bacteriology,* 171**,** 6197-6205.

LOMAN, N. J., CONSTANTINIDOU, C., CHRISTNER, M., ROHDE, H., CHAN, J. Z.-M., QUICK, J., WEIR, J. C., QUINCE, C., SMITH, G. P. & BETLEY, J. R. 2013. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104: H4. *The Jounral of the American Medical Association,* 309**,** 1502-1510.

LOMAN, N. J. & PALLEN, M. J. 2015. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology,* 13**,** 787.

LOVE, D. & LOVE, R. 1979. Pathology of proliferative haemorrhagic enteropathy in pigs. *Veterinary Pathology,* 16**,** 41-48.

LOVE, R., LOVE, D. & EDWARDS, M. 1977. Proliferative haemorrhagic enteropathy in pigs. *The Veterinary Record,* 100**,** 65-68.

LUMPPIO, H. L., SHENVI, N. V., SUMMERS, A. O., VOORDOUW, G. & KURTZ, D. M. 2001. Rubrerythrin and Rubredoxin Oxidoreductase in *Desulfovibrio vulgaris*: a Novel Oxidative Stress Protection System. *Journal of Bacteriology,* 183**,** 101-108.

LUO, C., KNIGHT, R., SILJANDER, H., KNIP, M., XAVIER, R. J. & GEVERS, D. 2015. ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology,* 33**,** 1045.

MACINTYRE, N., SMITH, D., SHAW, D., THOMSON, J. & RHIND, S. 2003a. Immunopathogenesis of experimentally induced proliferative enteropathy in pigs. *Veterinary Pathology,* 40**,** 421-432.

MACINTYRE, N., SMITH, D., SHAW, D., THOMSON, J. & RHIND, S. 2003b. Immunopathogenesis of experimentally induced proliferative enteropathy in pigs. *Veterinary Pathology Online,* 40**,** 421-432.

MACKINNON, M. J., GANDON, S. & READ, A. F. 2008. Virulence evolution in response to vaccination: the case of malaria. *Vaccine,* 26**,** C42-C52.

MAIDEN, M. C., BYGRAVES, J. A., FEIL, E., MORELLI, G., RUSSELL, J. E., URWIN, R., ZHANG, Q., ZHOU, J., ZURTH, K. & CAUGANT, D. A. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences,* 95**,** 3140-3145.

MAMIROVA, L., POPADIN, K. & GELFAND, M. S. 2007. Purifying selection in mitochondria, free-living and obligate intracellular proteobacteria. *BMC Evolutionary Biology,* 7**,** 17.

MARTTINEN, P., HANAGE, W. P., CROUCHER, N. J., CONNOR, T. R., HARRIS, S. R., BENTLEY, S. D. & CORANDER, J. 2011. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research,* 40**,** e6-e6.

MAU, B., GLASNER, J. D., DARLING, A. E. & PERNA, N. T. 2006. Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. *Genome Biology,* 7**,** R44.

MCADAM, P. R., RICHARDSON, E. J. & FITZGERALD, J. R. 2014. High-throughput sequencing for the study of bacterial pathogen biology. *Current Opinion in Microbiology,* 19**,** 106-113.

MCCLUSKEY, J., HANNIGAN, J., HARRIS, J. D., WREN, B. & SMITH, D. G. 2002. LsaA, an antigen involved in cell attachment and invasion, is expressed by *Lawsonia intracellularis* during infection in vitro and in vivo. *Infection and Immunity,* 70**,** 2899-2907.

MCINERNEY, J. O., MCNALLY, A. & O'CONNELL, M. J. 2017. Why prokaryotes have pangenomes. *Nature Microbiology,* 2**,** 17040.

MCORIST, S., BOID, R., LAWSON, G. & MCCONNELL, I. 1987. Monoclonal antibodies to intracellular campylobacter-like organisms of the porcine proliferative enteropathies. *Veterinary Record,* 121**,** 421-422.

MCORIST, S. & GEBHART, C. J. 2012. Proliferative enteropathy. *Diseases of Swine. 10th ed. Hoboken, New Jersey: Wiley-Blackwell Publishing***,** 811-819.

MCORIST, S., JASNI, S., MACKIE, R., BERSCHNEIDER, H., ROWLAND, A. & LAWSON, G. 1995. Entry of the bacterium ileal symbiont intracellularis into cultured enterocytes and its subsequent release. *Research in Veterinary Science,* 59**,** 255-260.

MCORIST, S., JASNI, S., MACKIE, R., MACINTYRE, N., NEEF, N. & LAWSON, G. 1993a. Reproduction of porcine proliferative enteropathy with pure cultures of ileal symbiont intracellularis. *Infection and Immunity,* 61**,** 4286-4292.

MCORIST, S., JASNI, S., MACKIE, R. A., MACINTYRE, N., NEEF, N. & LAWSON, G. H. 1993b. Reproduction of porcine proliferative enteropathy with pure cultures of ileal symbiont intracellularis. *Infection and Immunity,* 61.

MCORIST, S., MACINTYRE, N., STOKES, C. & LAWSON, G. 1992. Immunocytological responses in porcine proliferative enteropathies. *Infection and Immunity,* 60**,** 4184-4191.

MCORIST, S., ROBERTS, L., JASNI, S., ROWLAND, A., LAWSON, G., GEBHART, C. & BOSWORTH, B. 1996. Developed and resolving lesions in porcine proliferative enteropathy: possible pathogenetic mechanisms. *Journal of Comparative Pathology,* 115**,** 35-45.

MCORIST, S. & SMITS, R. 2007. Field evaluation of an oral attenuated *Lawsonia intracellularis* vaccine for porcine proliferative. *The Veterinary Record,* 161**,** 26-28.

METHÉ, B. A., NELSON, K. E., POP, M., CREASY, H. H., GIGLIO, M. G., HUTTENHOWER, C., GEVERS, D., PETROSINO, J. F., ABUBUCKER, S. & BADGER, J. H. 2012. A framework for human microbiome research. *Nature,* 486**,** 215.

MIRAJKAR, N. S., KELLEY, M. R. & GEBHART, C. J. 2017. Draft Genome Sequence of *Lawsonia intracellularis* Strain E40504, Isolated from a Horse Diagnosed with Equine Proliferative Enteropathy. *Genome Announcements,* 5**,** e00330-17.

MONTEIRO, L., BONNEMAISON, D., VEKRIS, A., PETRY, K. G., BONNET, J., VIDAL, R., CABRITA, J. & MÉGRAUD, F. 1997. Complex polysaccharides as PCR inhibitors in feces: *Helicobacter pylori* model. *Journal of Clinical Microbiology,* 35**,** 995-998.

MORAN, N. A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences,* 93**,** 2873-2878.

MORAN, N. A., MCLAUGHLIN, H. J. & SOREK, R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science,* 323**,** 379-382.

MORELLI, G., SONG, Y., MAZZONI, C. J., EPPINGER, M., ROUMAGNAC, P., WAGNER, D. M., FELDKAMP, M., KUSECEK, B., VOGLER, A. J. & LI, Y. 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics,* 42**,** 1140.

MOYA, A., PERETÓ, J., GIL, R. & LATORRE, A. 2008. Learning how to live together: genomic insights into prokaryote–animal symbioses. *Nature Reviews Genetics,* 9**,** 218.

NARRA, H. P. & OCHMAN, H. 2006. Of what use is sex to bacteria? *Current Biology,* 16**,** R705-R710.

NAVON-VENEZIA, S., KONDRATYEVA, K. & CARATTOLI, A. 2017. *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiology Reviews,* 41**,** 252-275.

NGUYEN, L.-T., SCHMIDT, H. A., VON HAESELER, A. & MINH, B. Q. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution,* 32**,** 268-274.

NURK, S., MELESHKO, D., KOROBEYNIKOV, A. & PEVZNER, P. A. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Research,* 27**,** 824-834.

OCTAVIA, S., MAHARJAN, R. P., SINTCHENKO, V., STEVENSON, G., REEVES, P. R., GILBERT, G. L. & LAN, R. 2010. Insight into evolution of *Bordetella pertussis* from comparative genomic analysis: evidence of vaccine-driven selection. *Molecular Biology and Evolution,* 28**,** 707-715.

OH, Y.-S., LEE, J.-B. & MCORIST, S. 2010. Microarray analysis of differential expression of cell cycle and cell differentiation genes in cells infected with *Lawsonia intracellularis*. *The Veterinary Journal,* 184**,** 340-345.

OLSON, N. D., TREANGEN, T. J., HILL, C. M., CEPEDA-ESPINOZA, V., GHURYE, J., KOREN, S. & POP, M. 2017. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics*.

PAGE, A. J., CUMMINS, C. A., HUNT, M., WONG, V. K., REUTER, S., HOLDEN, M. T., FOOKES, M., FALUSH, D., KEANE, J. A. & PARKHILL, J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics,* 31**,** 3691-3693.

PARADIS, M.-A., GEBHART, C. J., TOOLE, D., VESSIE, G., WINKELMAN, N. L., BAUER, S. A., WILSON, J. B. & MCCLURE, C. A. 2012. Subclinical ileitis: Diagnostic and performance parameters in a multi-dose mucosal homogenate challenge model. *Journal of Swine Health and Production,* 20**,** 137-141.

PARK, S., LEE, J.-B., KIM, K.-J., OH, Y.-S., KIM, M.-O., OH, Y.-R., HWANG, M.-A., LEE, J.-A. & LEE, S.-W. 2013. Efficacy of a commercial live attenuated *Lawsonia intracellularis* vaccine in a large scale field trial in Korea. *Clinical and Experimental Vaccine Research,* 2**,** 135-139.

PARK, S., WON, G., KIM, J., KIM, H. B. & LEE, J. H. 2018. Potent O-antigen-deficient (rough) mutants of *Salmonella Typhimurium* secreting *Lawsonia intracellularis* antigens enhance immunogenicity and provide single-immunization protection against proliferative enteropathy and salmonellosis in a murine model. *Veterinary Research,* 49**,** 57.

PARKS, D. H., IMELFORT, M., SKENNERTON, C. T., HUGENHOLTZ, P. & TYSON, G. W. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research,* 25**,** 1043-1055.

PARKS, D. H., RINKE, C., CHUVOCHINA, M., CHAUMEIL, P.-A., WOODCROFT, B. J., EVANS, P. N., HUGENHOLTZ, P. & TYSON, G. W. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology,* 2**,** 1533.

PASCU, C., COSTINAR, L., MERNEA, I., TĂTAR, D. & HERMAN, V. 2015. Prevalence of *Lawsonia intracellularis* Infections in Pig Herds from the Western Romania. *Agriculture and Agricultural Science Procedia,* 6**,** 378-381.

PEARSON, H. E., TORIBIO, J.-A. L., LAPIDGE, S. J. & HERNÁNDEZ-JOVER, M. 2016. Evaluating the risk of pathogen transmission from wild animals to domestic pigs in Australia. *Preventive Veterinary Medicine,* 123**,** 39-51.

PEIPONEN, K. S., TIRKKONEN, B. T., JUNNILA, J. J. T. & HEINONEN, M. L. 2018. Effect of a live attenuated vaccine against *Lawsonia intracellularis* in weaned and finishing pig settings in Finland. *Acta Veterinaria Scandinavica,* 60**,** 18.

PENDLETON, K. M., ERB-DOWNWARD, J. R., BAO, Y., BRANTON, W. R., FALKOWSKI, N. R., NEWTON, D. W., HUFFNAGLE, G. B. & DICKSON, R. P. 2017. Rapid pathogen identification in bacterial pneumonia using real-time metagenomics. *American Journal of Respiratory and Critical Care Medicine,* 196**,** 1610-1612.

PENG, Y., LEUNG, H. C., YIU, S.-M. & CHIN, F. Y. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics,* 28**,** 1420-1428.

PERIWAL, V. & SCARIA, V. 2014. Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics,* 31**,** 1-9.

PETERSEN, T. N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods,* 8**,** 785.

PEVZNER, P. A., TANG, H. & WATERMAN, M. S. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences,* 98**,** 9748-9753.

POHLNER, J., HALTER, R., BEYREUTHER, K. & MEYER, T. F. 1987. Gene structure and extracellular secretion of Neisseria gonorrhoeae IgA protease. *Nature,* 325**,** 458.

PRICE, L. B., STEGGER, M., HASMAN, H., AZIZ, M., LARSEN, J., ANDERSEN, P. S., PEARSON, T., WATERS, A. E., FOSTER, J. T. & SCHUPP, J. 2012. *Staphylococcus aureus* CC398: host adaptation and emergence of methicillin resistance in livestock. *MBio,* 3**,** e00305-11.

PULLIAM, J. R., EPSTEIN, J. H., DUSHOFF, J., RAHMAN, S. A., BUNNING, M., JAMALUDDIN, A. A., HYATT, A. D., FIELD, H. E., DOBSON, A. P. & DASZAK, P. 2011. Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis. *Journal of the Royal Society Interface***,** rsif20110223.

PUSTERLA, N. & GEBHART, C. 2009. Equine proliferative enteropathy caused by *Lawsonia intracellularis*. *Equine Veterinary Education,* 21**,** 415-419.

PUSTERLA, N. & GEBHART, C. 2013. Equine proliferative enteropathy–a review of recent developments. *Equine Veterinary Journal,* 45**,** 403-409.

PUSTERLA, N., GEBHART, C. J., LAVOIE, J.-P. & DROLET, R. 2013. *Lawsonia intracellularis*. *Equine Infectious Diseases E-Book***,** 316.

PUSTERLA, N., MAPES, S. & GEBHART, C. 2012. Further investigation of exposure to *Lawsonia intracellularis* in wild and feral animals captured on horse properties with equine proliferative enteropathy. *The Veterinary Journal,* 194**,** 253-255.

PUSTERLA, N., MAPES, S., JOHNSON, C., SLOVIS, N., PAGE, A. & GEBHART, C. 2010. Comparison of feces versus rectal swabs for the molecular detection of *Lawsonia intracellularis* in foals with equine proliferative enteropathy. *Journal of Veterinary Diagnostic Investigation,* 22**,** 741-744.

PUSTERLA, N., MAPES, S., REJMANEK, D. & GEBHART, C. 2008. Detection of *Lawsonia intracellularis* by real-time PCR in the feces of free-living animals from equine farms with documented occurrence of equine proliferative enteropathy. *Journal of Wildlife Diseases,* 44.

QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., ZHANG, W., GUAN, Y. & SHEN, D. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature,* 490**,** 55.

QUINCE, C., WALKER, A. W., SIMPSON, J. T., LOMAN, N. J. & SEGATA, N. 2017. Shotgun metagenomics, from sampling to analysis. *Nature Bbiotechnology,* 35**,** 833.

RAMBAUT, A., LAM, T. T., MAX CARVALHO, L. & PYBUS, O. G. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution,* 2**,** vew007.

REDDY, T. B., THOMAS, A. D., STAMATIS, D., BERTSCH, J., ISBANDI, M., JANSSON, J., MALLAJOSYULA, J., PAGANI, I., LOBOS, E. A. & KYRPIDES, N. C. 2014. The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification. *Nucleic Acids Research,* 43**,** D1099-D1106.

RIBER, U., CORDES, H., BOUTRUP, T. S., JENSEN, T. K., HEEGAARD, P. M. & JUNGERSEN, G. 2011. Primary infection protects pigs against re-infection with

Lawsonia intracellularis in experimental challenge studies. *Veterinary Microbiology,* 149**,** 406-414.

RIBER, U., HEEGAARD, P. M., CORDES, H., STÅHL, M., JENSEN, T. K. & JUNGERSEN, G. 2015. Vaccination of pigs with attenuated *Lawsonia intracellularis* induced acute phase protein responses and primed cell-mediated immunity without reduction in bacterial shedding after challenge. *Vaccine,* 33**,** 156-162.

RICHARDSON, E. J., BACIGALUPE, R., HARRISON, E. M., WEINERT, L. A., LYCETT, S., VRIELING, M., ROBB, K., HOSKISSON, P. A., HOLDEN, M. T. & FEIL, E. J. 2018. Gene exchange drives the ecological success of a multi-host bacterial pathogen. *Nature Ecology & Evolution***,** 1.

RINKE, C., SCHWIENTEK, P., SCZYRBA, A., IVANOVA, N. N., ANDERSON, I. J., CHENG, J.-F., DARLING, A., MALFATTI, S., SWAN, B. K. & GIES, E. A. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature,* 499**,** 431.

ROERINK, F., MORGAN, C., KNETTER, S., PASSAT, M.-H., ARCHIBALD, A., AIT-ALI, T. & STRAIT, E. 2018. A novel inactivated vaccine against *Lawsonia intracellularis* induces rapid induction of humoral immunity, reduction of bacterial shedding and provides robust gut barrier function. *Vaccine,* 36**,** 1500-1508.

ROUMAGNAC, P., WEILL, F.-X., DOLECEK, C., BAKER, S., BRISSE, S., CHINH, N. T., LE, T. A. H., ACOSTA, C. J., FARRAR, J. & DOUGAN, G. 2006. Evolutionary history of *Salmonella typhi*. *Science,* 314**,** 1301-1304.

SAIT, M., AITCHISON, K., WHEELHOUSE, N., WILSON, K., LAINSON, F. A., LONGBOTTOM, D. & SMITH, D. G. 2013. Genome sequence of *Lawsonia intracellularis* strain N343, isolated from a sow with hemorrhagic proliferative enteropathy. *Genome Announcements,* 1**,** e00027-13.

SAMPIERI, F., ALLEN, A. L., PUSTERLA, N., VANNUCCI, F. A., ANTONOPOULOS, A. J., BALL, K. R., THOMPSON, J., DOWLING, P. M., HAMILTON, D. L. & GEBHART, C. J. 2013a. The rabbit as an infection model for equine proliferative enteropathy. *Canadian Journal of Veterinary Research,* 77**,** 110-119.

SAMPIERI, F., VANNUCCI, F. A., ALLEN, A. L., PUSTERLA, N., ANTONOPOULOS, A. J., BALL, K. R., THOMPSON, J., DOWLING, P. M., HAMILTON, D. L. & GEBHART, C. J. 2013b. Species-specificity of equine and porcine *Lawsonia intracellularis* isolates in laboratory animals. *Canadian Journal of Veterinary Research,* 77**,** 261-272.

SCHLABERG, R., CHIU, C. Y., MILLER, S., PROCOP, G. W., WEINSTOCK, G., COMMITTEE, P. P., MICROBIOLOGY, C. O. L. P. O. T. A. S. F. & PATHOLOGISTS, M. R. C. O. T. C. O. A. 2017. Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Archives of Pathology and Laboratory Medicine,* 141**,** 776-786.

SCHMITZ-ESSER, S., HAFERKAMP, I., KNAB, S., PENZ, T., AST, M., KOHL, C., WAGNER, M. & HORN, M. 2008. *Lawsonia intracellularis* contains a gene encoding a functional rickettsia-like ATP/ADP translocase for host exploitation. *Journal of Bacteriology,* 190**,** 5746-5752.

SCHNURRBUSH, U., GRANDJOT, G., BREDE, W. 2014. The real pig handbook.

SCHOULS, L. M., SPALBURG, E. C., VAN LUIT, M., HUIJSDENS, X. W., PLUISTER, G. N., VAN SANTEN-VERHEUVEL, M. G., VAN DER HEIDE, H. G., GRUNDMANN, H., HECK, M. E. & DE NEELING, A. J. 2009. Multiple-locus variable number tandem repeat analysis of *Staphylococcus aureus*: comparison with pulsed-field gel electrophoresis and spa-typing. *PloS one,* 4**,** e5082.

SCHRADER, C., SCHIELKE, A., ELLERBROEK, L. & JOHNE, R. 2012. PCR inhibitors–occurrence, properties and removal. *Journal of Applied Microbiology,* 113**,** 1014-1026.

SCZYRBA, A., HOFMANN, P., BELMANN, P., KOSLICKI, D., JANSSEN, S., DRÖGE, J., GREGOR, I., MAJDA, S., FIEDLER, J. & DAHMS, E. 2017. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature Methods,* 14**,** 1063.

SEEMANN, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* 30**,** 2068-2069.

SEEMANN, T. 2015. snippy: fast bacterial variant calling from NGS reads.

SERRUTO, D., ADU-BOBIE, J., SCARSELLI, M., VEGGI, D., PIZZA, M., RAPPUOLI, R. & ARICÒ, B. 2003. *Neisseria meningitidis* App, a new adhesin with autocatalytic serine protease activity. *Molecular Microbiology,* 48**,** 323-334.

SHAO, L., KAMALU, O. & MAYER, L. 2005. Non-classical MHC class I molecules on intestinal epithelial cells: mediators of mucosal crosstalk. *Immunological Reviews,* 206**,** 160-176.

SHAPIRO, B. J. 2016. How clonal are bacteria over time? *Current Opinion in Microbiology,* 31**,** 116-123.

SHIMIZU, C., SHIBAHARA, T., TAKAI, S., KASUYA, K., CHIKUBA, T., MURAKOSHI, N., KOBAYASHI, H. & KUBO, M. 2010. *Lawsonia intracellularis* and virulent Rhodococcus equi infection in a thoroughbred colt. *Journal of Comparative Pathology,* 143**,** 303-308.

SIMPSON, J. T., WONG, K., JACKMAN, S. D., SCHEIN, J. E., JONES, S. J. & BIROL, I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Research***,** gr. 089532.108.

SINGH, A., GOERING, R. V., SIMJEE, S., FOLEY, S. L. & ZERVOS, M. J. 2006. Application of molecular techniques to the study of hospital infection. *Clinical Microbiology Reviews,* 19**,** 512-530.

SINGH, S. P., SALAMON, H., LAHTI, C. J., FARID-MOYER, M. & SMALL, P. M. 1999. Use of pulsed-field gel electrophoresis for molecular epidemiologic and population genetic studies of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology,* 37**,** 1927-1931.

SMITH, J. M. & HAIGH, J. 1974. The hitch-hiking effect of a favourable gene. *Genetics Research,* 23**,** 23-35.

SMITH, S., MCORIST, S. & GREEN, L. 1998. Questionnaire survey of proliferative enteropathy on British pig farms. *The Veterinary Record,* 142**,** 690-693.

SMITH, S. H., WILSON, A. D., VAN ETTINGER, I., MACINTYRE, N., ARCHIBALD, A. L. & AIT-ALI, T. 2014. Down-regulation of mechanisms involved in cell transport and maintenance of mucosal integrity in pigs infected with *Lawsonia intracellularis*. *Veterinary Research,* 45**,** 55.

SOLDEN, L., LLOYD, K. & WRIGHTON, K. 2016. The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Current Opinion in Microbiology,* 31**,** 217-226.

SPRATT, B. G. 2004. Exploring the concept of clonality in bacteria. *Genomics, Proteomics, and Clinical Bacteriology.* Springer.

SPRATT, B. G. & MAIDEN, M. C. 1999. Bacterial population genetics, evolution and epidemiology. *Philosophical Transactions of the Royal Society of London B: Biological Sciences,* 354**,** 701-710.

STEGE, H., JENSEN, T. K., MØLLER, K., BAEKBO, P. & JORSAL, S. 2000. Prevalence of intestinal pathogens in Danish finishing pig herds. *Preventive Veterinary Medicine,* 46**,** 279-292.

STEWART, R. D., AUFFRET, M. D., WARR, A., WISER, A. H., PRESS, M. O., LANGFORD, K. W., LIACHKO, I., SNELLING, T. J., DEWHURST, R. J. & WALKER, A. W. 2018. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Ccommunications,* 9**,** 870.

STILLS, H. F. 1991. Isolation of an intracellular bacterium from hamsters (*Mesocricetus auratus*) with proliferative ileitis and reproduction of the disease with a pure culture. *Infection and Immunity,* 59.

SUERBAUM, S. & ACHTMAN, M. 2004. *Helicobacter pylori*: recombination, population structure and human migrations. *International Journal of Medical Microbiology,* 294**,** 133-139.

SWAMINATHAN, B., BARRETT, T. J., HUNTER, S. B., TAUXE, R. V. & FORCE, C. P. T. 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging Infectious Diseases,* 7**,** 382.

TAKUNO, S., KADO, T., SUGINO, R. P., NAKHLEH, L. & INNAN, H. 2011. Population genomics in bacteria: a case study of *Staphylococcus aureus*. *Molecular Biology and Evolution,* 29**,** 797-809.

TAYLOR-BROWN, A., SPANG, L., BOREL, N. & POLKINGHORNE, A. 2017. Culture-independent metagenomics supports discovery of uncultivable bacteria within the genus *Chlamydia*. *Scientific Reports,* 7**,** 10661.

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L. & DURKIN, A. S. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences,* 102**,** 13950-13955.

TIBAYRENC, M. & AYALA, F. 2017. Is predominant clonal evolution a common evolutionary adaptation to parasitism in pathogenic parasitic protozoa, fungi, bacteria, and viruses? *Advances in Parasitology.* Elsevier.

TOFT, C. & ANDERSSON, S. G. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature Reviews Genetics,* 11**,** 465.

TREANGEN, T. J., ONDOV, B. D., KOREN, S. & PHILLIPPY, A. M. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology,* 15**,** 524.

TREANGEN, T. J. & SALZBERG, S. L. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics,* 13**,** 36.

TRUONG, D. T., FRANZOSA, E. A., TICKLE, T. L., SCHOLZ, M., WEINGART, G., PASOLLI, E., TETT, A., HUTTENHOWER, C. & SEGATA, N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods,* 12**,** 902.

TRUONG, D. T., TETT, A., PASOLLI, E., HUTTENHOWER, C. & SEGATA, N. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research*.

TSANG, A. K., LEE, H. H., YIU, S.-M., LAU, S. K. & WOO, P. C. 2017. Failure of phylogeny inferred from multilocus sequence typing to represent bacterial phylogeny. *Scientific Reports,* 7**,** 4536.

TURNER, D. P., WOOLDRIDGE, K. G. & ALA'ALDEEN, D. A. 2002. Autotransported serine protease A of *Neisseria meningitidis*: an immunogenic, surface-exposed outer membrane, and secreted protein. *Infection and Immunity,* 70**,** 4447-4461.

TYSON, G. W., CHAPMAN, J., HUGENHOLTZ, P., ALLEN, E. E., RAM, R. J., RICHARDSON, P. M., SOLOVYEV, V. V., RUBIN, E. M., ROKHSAR, D. S. & BANFIELD, J. F. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature,* 428**,** 37.

VAN BOVEN, M., MOOI, F. R., SCHELLEKENS, J. F., DE MELKER, H. E. & KRETZSCHMAR, M. 2005. Pathogen adaptation under imperfect vaccination: implications for pertussis. *Proceedings of the Royal Society of London B: Biological Sciences,* 272**,** 1617-1624.

VAN DEN BERG, B. 2010. Crystal structure of a full-length autotransporter. *Journal of Molecular Biology,* 396**,** 627-633.

VAN DEN WOLLENBERG, L., BUTLER, C., HOUWERS, D., DE GROOTV, M., PANHUIJZEN, H., VAN MAANEN, C. & VAN OLDRUITENBORGH-OOSTERBAAN, M. 2011. *Lawsonia intracellularis*-associated proliferative enteritis in weanling foals in the Netherlands. *Tijdschr Diergeneeskd,* 136**,** 565-570.

VAN DER HEIJDEN, H., BAKKER, J., ELBERS, A., VOS, J., WEYNS, A., DE SMET, M. & MCORIST, S. 2004. Prevalence of exposure and infection of Lawsonia intracellularis among slaughter-age pigs. *Research in veterinary science,* 77**,** 197-202.

VANNUCCI, F. & GEBHART, C. 2014. Recent advances in understanding the pathogenesis of *Lawsonia intracellularis* infections. *Veterinary Pathology,* 51**,** 465-477.

VANNUCCI, F. A., BECKLER, D., PUSTERLA, N., MAPES, S. M. & GEBHART, C. J. 2013a. Attenuation of virulence of *Lawsonia intracellularis* after in vitro passages and its effects on the experimental reproduction of porcine proliferative enteropathy. *Veterinary Microbiology,* 162.

VANNUCCI, F. A., BORGES, E. L., DE OLIVEIRA, J. S. V. & GUEDES, R. M. C. 2010. Intestinal absorption and histomorphometry of Syrian hamsters (*Mesocricetus auratus*) experimentally infected with *Lawsonia intracellularis*. *Veterinary Microbiology,* 145**,** 286-291.

VANNUCCI, F. A., FOSTER, D. & GEBHART, C. J. 2013b. Laser microdissection coupled with RNA-seq analysis of porcine enterocytesinfected with an obligate intracellular pathogen (*Lawsonia intracellularis*). *BMC Genomics,* 14.

VANNUCCI, F. A., FOSTER, D. N. & GEBHART, C. J. 2012a. Comparative transcriptional analysis of homologous pathogenic and non-pathogenic *Lawsonia intracellularis* isolates in infected porcine cells. *PLoS One,* 7.

VANNUCCI, F. A., FOSTER, D. N. & GEBHART, C. J. 2013c. Laser microdissection coupled with RNA-seq analysis of porcine enterocytes infected with an obligate intracellular pathogen (*Lawsonia intracellularis*). *BMC Genomics,* 14**,** 421.

VANNUCCI, F. A., KELLEY, M. R. & GEBHART, C. J. 2013d. Comparative genome sequencing identifies a prophage-associated genomic island linked to host adaptation of *Lawsonia intracellularis* infections. *Veterinary Research,* 44**,** 49.

VANNUCCI, F. A., PUSTERLA, N., MAPES, S. M. & GEBHART, C. 2012b. Evidence of host adaptation in *Lawsonia intracellularis* infections. *Veterinary Research,* 43.

VANNUCCI, F. A., PUSTERLA, N., MAPES, S. M. & GEBHART, C. 2012c. Evidence of host adaptation in *Lawsonia intracellularis infections*. *Veterinary Research,* 43**,** 53.

VANNUCCI, F. A., WATTANAPHANSAK, S. & GEBHART, C. J. 2012d. An alternative method for cultivation of *Lawsonia intracellularis*. *Journal of Clinical Microbiology***,** JCM. 05976-11.

VERBRUGGHE, E., BOYEN, F., VAN PARYS, A., VAN DEUN, K., CROUBELS, S., THOMPSON, A., SHEARER, N., LEYMAN, B., HAESEBROUCK, F. & PASMANS, F. 2011. Stress induced *Salmonella Typhimurium* recrudescence in pigs coincides with cortisol induced increased intracellular proliferation in macrophages. *Veterinary Research,* 42**,** 118.

VLISIDOU, I., LYTE, M., VAN DIEMEN, P. M., HAWES, P., MONAGHAN, P., WALLIS, T. S. & STEVENS, M. P. 2004. The neuroendocrine stress hormone norepinephrine augments *Escherichia coli* O157: H7-induced enteritis and adherence in a bovine ligated ileal loop model of infection. *Infection and Immunity,* 72**,** 5446-5451.

VOLLMERS, J., WIEGAND, S. & KASTER, A.-K. 2017. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective-Not only size matters! *PloS one,* 12**,** e0169662.

WALTER, D., GEBHART, C., KROLL, J., HOLCK, J. & CHITTICK, W. 2004. Serologic profiling and vaccination timing for *Lawsonia intracellularis*. *Journal of Swine Health and Production,* 12**,** 310-313.

WATSON, E., CLARK, E. M., ALBERDI, M. P., INGLIS, N. F., PORTER, M., IMRIE, L., MCLEAN, K., MANSON, E., LAINSON, A. & SMITH, D. G. 2011. A novel *Lawsonia intracellularis* autotransporter protein is a prominent antigen. *Clinical and Vaccine Immunology***,** CVI. 05073-11.

WATTANAPHANSAK, S., GEBHART, C. J., ANDERSON, J. M. & SINGER, R. S. 2010. Development of a polymerase chain reaction assay for quantification of *Lawsonia intracellularis*. *Journal of Veterinary Diagnostic Investigation,* 22**,** 598-602.

WEINERT, L. A., CHAUDHURI, R. R., WANG, J., PETERS, S. E., CORANDER, J., JOMBART, T., BAIG, A., HOWELL, K. J., VEHKALA, M. & VÄLIMÄKI, N. 2015. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nature Communications,* 6**,** 6740.

WELLS, T. J., TREE, J. J., ULETT, G. C. & SCHEMBRI, M. A. 2007. Autotransporter proteins: novel targets at the bacterial cell surface. *FEMS Microbiology Letters,* 274**,** 163-172.

WILSON, M. R., SHANBHAG, N. M., REID, M. J., SINGHAL, N. S., GELFAND, J. M., SAMPLE, H. A., BENKLI, B., O'DONOVAN, B. D., ALI, I. K. & KEATING, M. K. 2015. Diagnosing *Balamuthia mandrillaris E ncephalitis* With Metagenomic Deep Sequencing. *Annals of Neurology,* 78**,** 722-730.

WON, G. & LEE, J. H. 2018. Antigenic and functional profiles of a *Lawsonia intracellularis* protein that shows a flagellin-like trait and its immuno-stimulatory assessment. *Veterinary Research,* 49**,** 17.

WOOD, D. E. & SALZBERG, S. L. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology,* 15**,** R46.

WU, Z., LING, Y., TIAN, D., PAN, Q., HEEGAARD, P. M. & HE, C. 2014. Seroprevalence of *Lawsonia intracellularis* antibodies in intensive pig farms in China. *BMC Veterinary Research,* 10**,** 100.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution,* 24**,** 1586-1591.

ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Research,* 18.

Appendix

Supplementary Tables

Supplementary Table 1 SNPs identified from pairwise comparison between Enterisol® Ileitis and 8163

Location	8163	ENT	Evidence	Type	Effect	Predicted gene	Putative protein product
Chr	T	G	G:353 T:13	CDS	NS	ros	Transcriptional regulatory protein ros
Chr	C	T	T:512 C:0	CDS	S	purN	Phosphoribosylglycinamide formyltransferase
Chr	A	G	G:444 A:0	CDS	NS	msrB	Peptide methionine sulfoxide reductase MsrB
Chr	A	T	T:479 A:1	CDS	stop_lost		hypothetical protein
Chr	T	C	C:479 T:0	CDS	NS		hypothetical protein
Chr	T	C	C:449 T:1	CDS	S		hypothetical protein
Chr	C	T	T:498 C:0				
Chr	G	A	A:365 G:1	CDS	NS	coaBC	Coenzyme A biosynthesis bifunctional protein CoaBC
Chr	T	A	A:222 T:5	CDS	NS	zupT	Zinc transporter ZupT
Chr	G	A	A:508 G:0	CDS	S		Gluconeogenesis factor
Chr	A	G	G:411 A:0	CDS	NS	rne	Ribonuclease E
Chr	A	G	G:446 A:0	CDS	S	mepM_3	Murein DD-endopeptidase MepM
Chr	G	C	C:519 G:0	CDS	NS	parB	putative chromosome-partitioning protein ParB
Chr	G	A	A:495 G:0	CDS	NS		putative zinc protease
Chr	A	G	G:418 A:0				
Chr	T	C	C:415 T:0	CDS	S	pnp	Polyribonucleotide nucleotidyltransferase
Chr	T	C	C:413 T:0	CDS	S	bamD_2	Outer membrane protein assembly factor BamD
Chr	A	G	G:432 A:0	CDS	NS	murE	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2, 6-diaminopimelate ligase

ENT = Enterisol® Ileitis, CDS = Coding Sequence, NS = non-synonymous, S = synonymous

Supplementary Table 1 SNPs identified from pairwise comparison between Enterisol® Ileitis and 8163 (continue)

Location	Ref	Alt	Evidence	Type	Effect	Predicted gene	Putative protein product
Chr	A	G	G:414 A:12	CDS	NS	cheR	Chemotaxis protein methyltransferase
Plasmid 1	C	T	T:490 C:0	CDS	S	gmhA1	Phosphoheptose isomerase 1
Plasmid 3	T	C	C:503 T:1				
Plasmid 3	T	C	C:476 T:0	CDS	NS		hypothetical protein
Plasmid 3	G	A	A:406 G:1	CDS	S	yabJ	2-iminobutanoate/2-iminopropanoate deaminase
Plasmid 3	G	A	A:421 G:0	CDS	NS		hypothetical protein
Plasmid 3	T	C	C:476 T:0	CDS	NS		hypothetical protein

ENT = Enterisol® Ileitis, CDS = Coding Sequence, NS = non-synonymous, S = synonymous

Supplementary Table 2 Annotation of genes within elevated SNP density region identified by Gubbins

Node	Start	End	Length (bp)	PHE/MN1-00 locus tag	Predicted gene name	Putative protein product		COG
Node_31	54844	58949	4105	LI_RS00230		Hypothetical protein		
				LI_RS00235		Hypothetical protein		
	67025	72039	5014	LI_RS00275		Hypothetical protein		
				LI_RS00280		ATP-dependent exoDNAse subunit beta		L
				LI_RS00285		diphosphate--fructose-6-phosphate phosphotransferase	1	G
	168774	171154	2380	LI_RS00745		Hypothetical protein		
				LI_RS00750		Hypothetical protein		
				LI_RS00755		Hypothetical protein		
				LI_RS00760		Glycosyltransferase		M
	262539	266911	4372	LI_RS01255		SsrA-binding protein		O
				LI_RS01260		heme ABC transporter ATP-binding protein		S
				LI_RS01265		malonyl CoA-ACP transacylase		I
				LI_RS01270		fructose 1,6-bisphosphatase		G
				LI_RS01275	*argS*	arginine--tRNA ligase		J
	337235	341640	4405	LI_RS01575		Hypothetical protein		
				LI_RS01580		Hypothetical protein		
				LI_RS01585		ATPase AAA		L
	645192	647451	2259	LI_RS02965		Hypothetical protein		
				LI_RS02970		Hypothetical protein		
				LI_RS02980		Hypothetical protein		

Supplementary Table 2. Annotation of genes within elevated SNP density region identified by Gubbins (continue)

Node	Start	End	Length (bp)	PHE/MN1-00 locus tag	Predicted gene name	Putative protein product	COG
				LI_RS02985		Hypothetical protein	
	736940	742494	5554	LI_RS03370	*radA*	DNA repair protein RadA	O
				LI_RS03375	*clpS*	ATP-dependent Clp protease adapter protein ClpS	O
				LI_RS03380	*clpA*	ATP-dependent Clp protease ATP-binding subunit ClpA	O
				LI_RS03385	*aat*	Leucyl/phenylalanyl-tRNA--protein transferase	O
	854925	859086	4161	LI_RS03845	*uvrC*	UvrABC system protein C	L
				LI_RS03850	*maf*	MAF-like protein	D
				LI_RS03855		Hypothetical protein	
				LI_RS03860	*asnS*	asparaginyl-tRNA ligase	J
	884949	887056	2107	LI_RS03970		histidinol phosphate phosphatase	E
				LI_RS03975		S-adenosylmethionine synthase	H
	901727	904230	2503	LI_RS04065	*flgH*	Flagellar L-ring protein	N
				LI_RS04070	*flgI*	Flagellar P-ring protein	N
				LI_RS04075	*flgJ*	Peptidoglycan hydrolase FlgJ	N
				LI_RS04080		Hypothetical protein	
	1016094	1020028	3934	LI_RS04580		exodeoxyribonuclease III	L
				LI_RS04585		Hypothetical protein	
				LI_RS04590		invasin	M
				LI_RS04600	*katE*	hydroperoxidase II	P

Supplementary Table 2. Annotation of genes within elevated SNP density region identified by Gubbins (continue)

Node	Start	End	Length (bp)	PHE/MN1-00 locus tag	Predicted gene name	Putative protein product	COG
	1028994	1030611	1617	LI_RS04620		glycosyl transferase	M
				LI_RS04625		Hypothetical protein	
				LI_RS04630	*trmFO*	methylenetetrahydrofolate--tRNA-(uracil-5-)-methyltransferase TrmFO	J
	1052502	1054328	1826	LI_RS04735	*rluB*	Ribosomal large subunit pseudouridine synthase B	J
				LI_RS04740		ABC transporter ATP-binding protein	
	1082064	1084437	2373	LI_RS04855		Hypothetical protein	
	1106476	1108477	2001	LI_RS04935		glutamate-1-semialdehyde 2,1-aminomutase	H
				LI_RS04940		Hypothetical protein	
	1304615	1306271	1656	LI_RS05930	*feoB*	Fe(2+) transporter FeoB	P
	1320064	1322822	2758	LI_RS05990		Pyruvate kinase	G
				LI_RS05995	*mraZ*	Transcriptional regulator MraZ	L
				LI_RS06000		Ribosomal RNA small subunit methyltransferase H	
	1401100	1402722	1622	LI_RS06385		5'-nucleotidase	F
	1458166	1459390	1224	LI_RS06660		Hypothetical protein	
	1526467	1530281	3814	LI_RS06910		Hypothetical protein	
				LI_RS06920		Hypothetical protein	

Supplementary Table 2. Annotation of genes within elevated SNP density region identified by Gubbins (continue)

Node	Start	End	Length (bp)	PHE/MN1-00 locus tag	Predicted gene name	Putative protein product	COG
	1558053	1563787	5734	LI_RS07000		Hypothetical protein	
				LI_RS07005		Hypothetical protein	
				LI_RS07010		Hypothetical protein	
				LI_RS07015		Hypothetical protein	
	1597204	1601029	3825	LI_RS07085		Hypothetical protein	
				LI_RS07090		Hypothetical protein	
	1610335	1612400	2065	LI_RS07115		Hypothetical protein	
	1614484	1620518	6034	LI_RS07125		Hypothetical protein	
				LI_RS07130		Hypothetical protein	
				LI_RS07135		Hypothetical protein	
				LI_RS07140		Hypothetical protein	
				LI_RS07145		Hypothetical protein	
				LI_RS07150		Hypothetical protein	
				LI_RS07155		Hypothetical protein	
				LI_RS07160		Hypothetical protein	
	1655893	1656562	669	LI_RS07245		Hypothetical protein	
	1678901	1681460	2559	LI_RS07310	*mcpQ*	Methyl-accepting chemotaxis protein McpQ	T

Supplementary Table 2. Annotation of genes within elevated SNP density region identified by Gubbins (continue)

Node	Start	End	Length (bp)	PHE/MN1-00 locus tag	Predicted gene name	Putative protein product	COG
E40504	67988	71926	3938	LI_RS00280		ATP-dependent exoDNAse subunit beta	L
				LI_RS00285		diphosphate--fructose-6-phosphate 1-phosphotransferase	G
	222059	226567	4508	LI_RS01065	*glmM*	Phosphoglucosamine mutase	G
				LI_RS01070		UTP-glucose-1-phosphate uridylyltransferase	M
				LI_RS01075		Primosomal protein N'	L
	612021	615254	3233	LI_RS02810		Octaprenyl diphosphate synthase	H
				LI_RS02815		phosphoribosylformylglycinamidine synthase	F
	775168	779510	4342	LI_RS03550	*valS*	Valine-tRNA ligase	J
				LI_RS03555		Hypothetical protein	
	850153	855379	5226	LI_RS03820		Rubrerythrin	C
				LI_RS03825		Rubredoxin	C
				LI_RS03830		heptosyltransferase ii	M
				LI_RS03835		Hypothetical protein	
				LI_RS03840		Hypothetical protein	
				LI_RS03845	*uvrC*	UvrABC system protein C	L
	1017187	1020508	3321	LI_RS04585		Hypothetical protein	
				LI_RS04590		Invasin	M
				LI_RS04600	*katE*	hydroperoxidase II	P

Supplementary Table 2. Annotation of genes within elevated SNP density region identified by Gubbins (continue)

Node	Start	End	Length (bp)	PHE/MN1-00 locus tag	Predicted gene name	Putative protein product	COG
	1045777	1050279	4502	LI_RS04710		cobalt transporter CbiM	P
				LI_RS04715		Hypothetical protein	
				LI_RS04720		membrane protein	M
				LI_RS04725		membrane metal-binding protein	M
	1280654	1283550	2896	LI_RS05820		Hypothetical protein	
				LI_RS05825		Fe-S-cluster-containing hydrogenase components 1	C
				LI_RS05830		dTDP-4-dehydrorhamnose reductase	M
				LI_RS05835		Hypothetical protein	M
	1384218	1390241	6023	LI_RS06300		Hypothetical protein	
				LI_RS06305		Hypothetical protein	
				LI_RS06310		regulatory protein PcrH	
				LI_RS06315		Hypothetical protein	
				LI_RS06320		Hypothetical protein	
	1471830	1475023	3193	LI_RS06705		Hypothetical protein	
				LI_RS06710		SAM-dependent methyltransferase	
				LI_RS06715		SAM-dependent methyltransferase	
	1505432	1512364	6932	LI_RS06830		Hypothetical protein	
				LI_RS06835		nucleotidyl transferase	M
				LI_RS06840		Galactokinase	
				LI_RS06845		inositol 2-dehydrogenase	
				LI_RS06850		Dehydrogenase	

Supplementary Table 2. Annotation of genes within elevated SNP density region identified by Gubbins (continue)

Node	Start	End	Length (bp)	PHE/MN1-00 locus tag	Predicted gene name	Putative protein product	COG
				LI_RS06855		UDP-glucose 6-dehydrogenase	
				LI_RS06860		Hypothetical protein	
	1534328	1537283	2955	LI_RS07005		Hypothetical protein	
				LI_RS06945		Hypothetical protein	
				LI_RS06950		Hypothetical protein	
				LI_RS06955		Hypothetical protein	
	1555472	1568411	12939	LI_RS06995		Hypothetical protein	
				LI_RS07000		Hypothetical protein	
				LI_RS07005		Hypothetical protein	
				LI_RS07015		Hypothetical protein	
				LI_RS07020		Hypothetical protein	
	1614484	1619372	4888	LI_RS07125		Hypothetical protein	
				LI_RS07130		Hypothetical protein	
				LI_RS07135		Hypothetical protein	
				LI_RS07140		Hypothetical protein	
				LI_RS07145		Hypothetical protein	
				LI_RS07150		Hypothetical protein	
				LI_RS07155		Hypothetical protein	
	1651684	1659717	8033	LI_RS07245		Hypothetical protein	
	1677820	1683500	5680	LI_RS07305		cytochrome B6	G

Supplementary Table 2. Annotation of genes within elevated SNP density region identified by Gubbins (continue)

Node	Start	End	Length (bp)	PHE/MN1-00 locus tag	Predicted gene name	Putative protein product	COG
				LI_RS07310	*mcpQ*	Methyl-accepting chemotaxis protein McpQ	T
				LI_RS07315		Hypothetical protein	
				LI_RS07320		Hypothetical protein	